A 0.3V VDDmin 4+2T SRAM for Searching and In-Memory Computing Using 55nm DDC Technology

Qing Dong[1], Supreet Jeloka[1], Mehdi Saligane[1], Yejoong Kim[1], Masaru Kawaminami[2], Akihiko Harada[2], Satoru Miyoshi[2], David Blaauw[1], Dennis Sylvester[1]
[1]University of Michigan, Ann Arbor, and [2]Fujitsu Semiconductor America, Inc. (email:qingdong@umich.edu)

Abstract

A 4+2T SRAM is proposed that offers searching and logic functions. The cell uses the N-well as the write wordline (WL) and eliminates the access transistors. Decoupled read paths enable reliable multi-word activation for in-memory Boolean logic functions. The SRAM can reconfigure to BCAM/TCAM for searching operations, with 0.13fJ/search/bit at 0.35V. Forty test chips in 55nm deeply depleted channel (DDC) technology achieve worst-case 0.3V VDDmin.

Introduction

Von Neumann architectures continuously transfer data between memory and computing elements, incurring energy and latency costs that can dominate system power and performance. To minimize this data movement overhead, in-memory computing allows for data processing inside on-chip memories [1, 2]. In-memory computing activates multiple rows simultaneously and computes results directly on the bitline (BL). Computation results are immediately available as the memory is accessed, saving clock cycles and interconnect energy.

Conventional 6T SRAM suffers from degraded read noise margin (RNM) when multiple rows are activated, limiting its application to in-memory computing and resulting in high VDDmin [2]. 8T SRAM improves RNM by decoupling read and write paths but incurs 30% area overhead or more [3]. We propose a 4+2T SRAM cell that uses the N-well as a write WL, eliminating the access transistors and resulting in a 4T-core memory cell. Two decoupled read paths (2T) significantly improve RNM, enabling reliable multi-word activation for logic operations and a low VDDmin while limiting area overhead to only 12% over commercial push-rule 6T SRAM in the same technology. Using dual sense amplifiers (SA), Boolean logic functions (AND, OR, XOR) between the two activated words can be realized. Furthermore, with separated RBL/RBLB and RWL/RWLB, the SRAM can be configured as a BCAM or TCAM, enabling search operations, while maintaining normal one-cycle write to store data instead of the two-cycle write in [2]. The memory cell is designed using pushed rules in 55nm DDC technology, which offers a high body coefficient and low process variation. Measurement of 40 dies, including corner wafers, show a worst-case 0.3V VDDmin for SRAM operation. The BCAM achieves 0.13fJ/search/bit at 0.35V.

Memory Cell and Write Method

Fig.1 shows the schematic of the proposed 4+2T SRAM cell. The cross-coupled inverters have separated VDD terminals, which serve as WBL and WBLB. Due to the strong body effect in DDC technology, the N-well can be used as WWL. Two decoupled read ports (2T) are used for read and logic operations. In CAM mode, RBL/RBLB and RWL/RWLB are configured as SL/SLB and ML/MLB, respectively. Fig. 1 also shows the layout and lithographic simulation of the proposed 4+2T cell using pushed rules. Cell area is $265F^2$. The WWL (N-well) runs horizontally, and RWL, RWLB, and GND contacts are shared with adjacent cells. The table in Fig. 2 summarizes the voltages applied on each terminal for memory, CAM, and logic operations.

Fig. 3 shows the write scheme applied to the 4T structure. In standby, both WBL and WBLB are set to VDD, and WWL is at a higher voltage VDDH. To write 0 into a storage node, the selected WBL is lowered from VDD to GND, while WBLB remains at VDD. Once WWL is asserted low, the selected PMOS device becomes much stronger due to its forward body bias. This will short WBL/WBLB with the internal cell node and write into the selected cell. However, there are two types of half-select disturbances to consider. Cells on the selected column also have lowered WBL, which can potentially flip them. Higher VDDH is thus applied to the WWL of these cells to weaken their PMOS devices and alleviate column disturbances. Conversely, all cells in the selected row have stronger PMOS devices during a write, increasing their chance of un-intended write. Since their WBL/WBLB are kept at VDD, increasing VDD reduces row disturbances. Fig. 3 shows the measured write margin of cell write, column disturbances, and row disturbances. The green region indicates >5σ write and disturbance margins. Column disturbance occurs at high VDD and low VDDH; row disturbance occurs at low VDD. Operating points centered in the green region (>5σ) have at least ± 200mV VDDH/VDD margin, ensuring robust write.

Differential Read and Logic Operations

Basic read operation is realized with a single decoupled read port similar to a 5T or 7T SRAM [4-5]. The proposed design uses a differential read to accelerate read speed and enable logic operations. During a normal read, one pair of RWL/RWLB is activated (pulled low), and one BL discharges while the other remains high (Fig. 4(a)). The two small column-wise SAs are connected in parallel to form a larger SA, accelerating read operations and reducing leakage current from unselected cells [4-5]. For logic operations (Fig. 4(b)), two pairs of WLs are activated simultaneously. RBL remains high only if both cell nodes (A and B) store 0, and RBL therefore represents the NAND of A and B. Similarly, RBLB is connected to the complementary nodes and provides the OR of A and B. Using two differential SAs, NAND/AND and NOR/OR is computed simultaneously and a NOR gate between the two SA outputs generates A XOR B. All Boolean logic functions are computed in a single read cycle. Since each SA is small, area overhead is <5% compared to a normal SRAM.

CAM Configurations

Fig. 5 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 match lines ML/MLB. For BCAM operation, ML and MLB in a row are shorted together as one matching line. If all the input data match the stored data, ML remains high; otherwise ML discharges. Each ML has a SA to evaluate the results, similar to a conventional BCAM [6]. Unlike a previous 6T BCAM [2] that requires transposed data storage and two cycles write, the proposed BCAM stores data in a normal row-wise fashion. Moreover, RNM is not degraded when multiple rows are activated, in contrast to [2]. 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. The searching and sensing method of TCAM is the same as in BCAM.

Measurement Results

The proposed SRAM was fabricated in 55nm DDC technology (Fig. 12). Array efficiency is 65% for a 128×128 pushed rule array including all peripherals. Fig. 6 shows the write frequency and energy across VDD and VDDH for a typical die. At 0.8V VDD, the write frequency is 600MHz. The minimum supply voltage is 0.25V/0.30V for VDD/VDDH, and optimal write energy is 4.02fJ/bit at VDD of 0.35V. For BCAM operation, VDDmin is ~0.35V and the optimal energy/search is 0.13fJ/bit (Fig. 7), which are a 2.1× and 3.2× improvement over [2], respectively. Fig. 8 shows the frequency and energy across VDD for normal reads and logic operations. For a typical die, read VDDmin is ~0.25V, whereas it is ~0.35V for logic operations since they employ single-port sensing and half-strength SAs. Optimal read energy is 3.96fJ/bit at 0.35V; energy at 0.25V is higher since the leakage energy overhead exceeds the reduction in dynamic energy. The logic frequency is 30% slower and energy/logic operation is 50% higher than a normal read operation. However the logic functions operate on 2 words simultaneously instead of a single word as in normal read. Therefore, the total latency (1.3 cycles) is 70% lower than a conventional 2-cycle read followed by logic. Also, energy is at least 50% lower than a 2-cycle read followed by logic. Fig. 9 shows the measured VDDmin across temperature for both read/write and hold. Hold VDDmin is ~0.2V at 25°C with 1.6μW leakage power (Fig. 10). Fig. 11 shows the within-wafer VDDmin distribution of 20 TT corner dies and the average VDDmin distribution of split wafers in each corner. The worst-case VDDmin is 0.3V among the 40 dies. Fig. 13 compares this work with other decoupled SRAM and CAM works.

References

[1] M. Kang et al., ISCAS 2015.
[2] S. Jeloka et al., JSSC Apr 2016.
[3] J. Keane et al., ISSCC 2016.
[4] D. Jeon, et al., VLSIC 2015.
[5] M. Chen et al., VLSIC 2012.
[6] A. Agarwal et al., ESSCIRC 2011.
[7] N. Verma et al., ISSCC 2007.

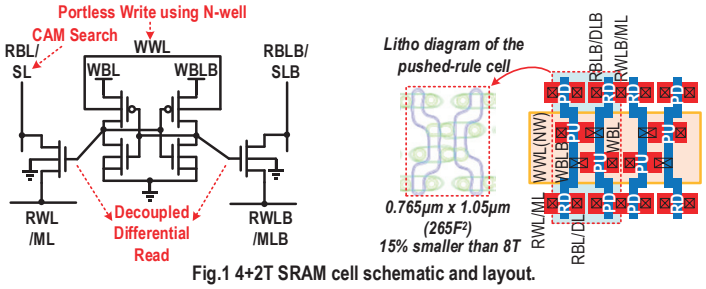
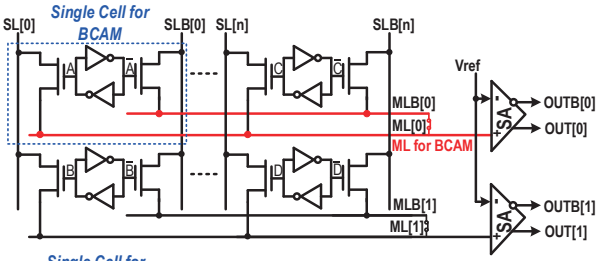


Fig.1 4+2T SRAM cell schematic and layout.

		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.2 Operation table.



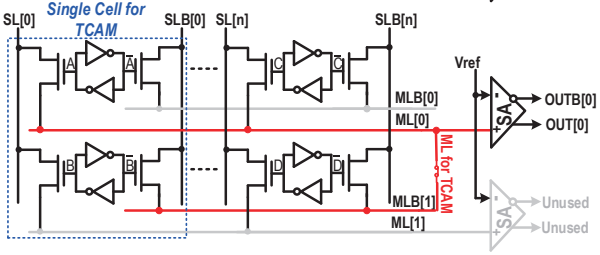


Fig.5 BCAM and TCAM configuration.

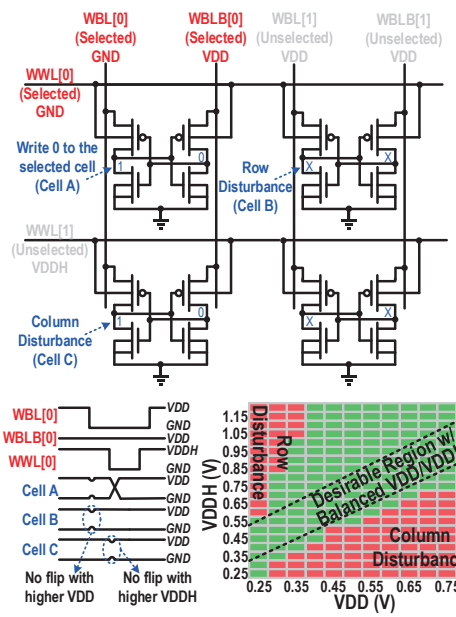


Fig.3 Write method and measured write Shmoo plot of 16kb array.

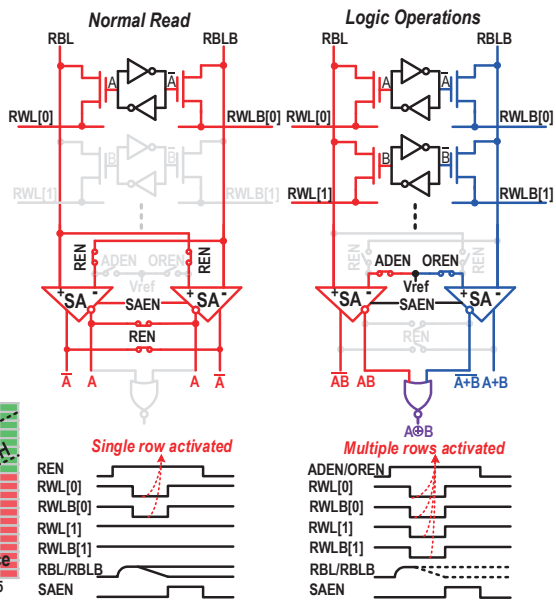


Fig.4 Comparison between normal read and Boolean logic operations (AND/OR/XOR).

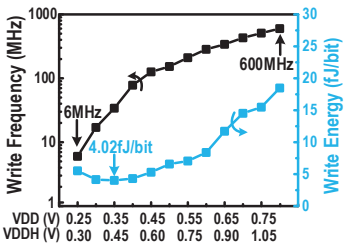


Fig.6 Write frequency and energy across VDD/VDDH.

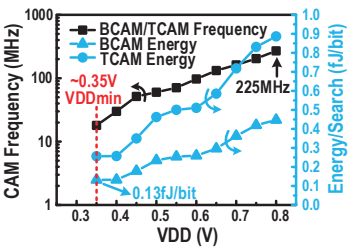
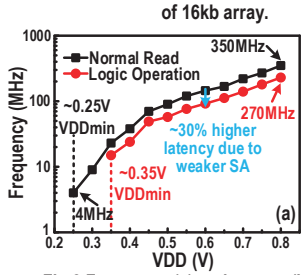


Fig.7 BCAM/TCAM frequency and energy across VDD.



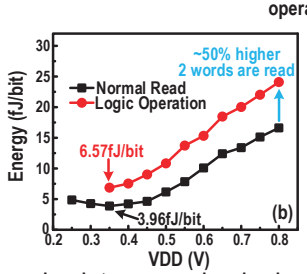


Fig.8 Frequency (a) and energy (b) comparison between normal read and logic operation.

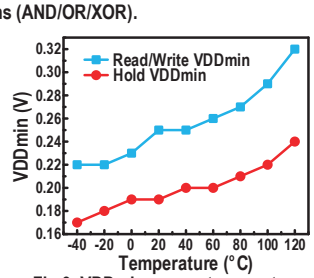


Fig.9. VDDmin across temperature.

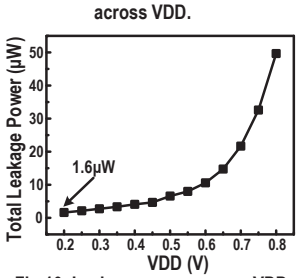


Fig 10. Leakage power across VDD.

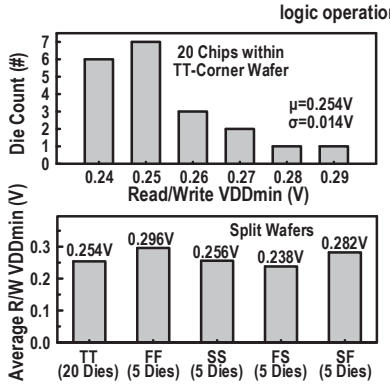


Fig.11 Within-wafer and split-wafer VDDmin distribution.

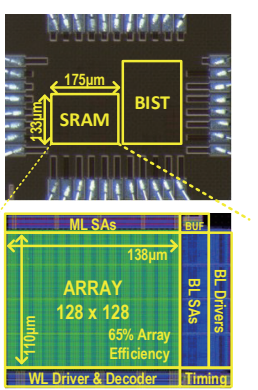


Fig.12 Die photo and block diagram.

		This work	Decoupled SRAM Work			CAM Work	
			[4]	[5]	[7]	[2]	[6]
Function		SRAM/CAM/Logic	SRAM	SRAM	SRAM	SRAM/CAM/Logic	BCAM
Technology		55nm DDC	40nm	65nm	65nm	28nm FDSOI	32nm
Cell Type		4+2T	5T	7T	8T	6T	11T
Cell Area Scaled to 6T		1.12x	0.93x	1.15x	1.3x	1x	>2x
Pushed-Rule Cell		YES	NO	NO	NO	YES	NO
Array Size		128 x 128 (16kb)	4Mb	256 x 128 (32kb)	256 x 128 x 8 (256kb)	64 x 64 (4kb)	64 x 64 x 4 (16kb)
Array Efficiency		65%	55%	46%	NA	60%	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)		
CAM VDDmin (V)		0.35				0.75	0.5
BCAM	Freq. (MHz)	270 (0.8V) 18 (0.35V)				370 (1V)	NA
	Energy (fJ/bit)[2]	0.45 (0.8V) 0.13 (0.35V)				0.6 (1V)	0.3 (0.5V)
Logic	Freq. (MHz)	230 (0.8V) 15 (0.35V)				594 (1V)	NA
	Energy (fJ/bit)[1]	24.1 (0.8V) 6.6 (0.35V)				NA	NA

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

Fig.13 Comparison table.