A Differential 2R Crosspoint RRAM Array with Zero Standby Current

Pi-Feng Chiu, *Student Member, IEEE*, and Borivoje Nikolić, *Senior Member, IEEE*
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA
pfchiu@eecs.berkeley.edu

***Abstract*— Memory power consumption dominates mobile system energy budgets in scaled technologies. Fast nonvolatile memories (NVMs) offer a tremendous opportunity to eliminate memory leakage current during standby mode. Resistive random access memory (RRAM) in a crosspoint structure is considered to be one of the most promising emerging NVMs. However, the absence of access transistors puts significant challenges on the write/read operation. In this paper, we propose a differential 2R crosspoint structure with array segmentation and sense-before-write techniques. A 64KB RRAM is constructed and simulated in a 28/32nm CMOS predictive technology model (PTM) and a Verilog-A RRAM model. This design offers an opportunity for using RRAM as a cache for increasing energy efficiency in mobile computing.**

***Keywords*— RRAM, memristor, nonvolatile memory, crosspoint, cache, zero standby current**

I. Introduction

An energy-efficient memory system is necessary for continued scaling of mobile systems into nanometer technologies. Mobile devices are idle more than 90% of the time, highlighting the need to minimize standby energy consumption. As the technology scaling trend continues, leakage current in SRAM-based cache memories will dominate energy consumption in standby mode. Nonvolatile memories can be powered down completely, eliminating the leakage current. Flash memory [1], the most popular nonvolatile memory, has a large storage density and small cell size. However, slow program/erase (P/E) speeds make it too impractical for caches, and physical limitations associated with oxide thickness prevent flash memory from continued scaling. Therefore, there is a perceived need for a high-speed nonvolatile memory that can be used as a universal memory, replacing both flash memory and SRAM.

New memory technologies include ferroelectric memory (FeRAM) [2], spin-transfer torque memory (STT-RAM) [3], phase-change memory (PRAM) [4], and resistive memory (RRAM) [5]. FeRAM has limited density due to scaling difficulties. PRAM is a thermally-driven process, which suffers from high programming current, low endurance and long-term resistance drift. STT-RAM has high endurance and high switching speed, and is being evaluated as a successor to DRAM. However, the resistance ratio between two states is low, which is a yield concern. RRAM is one of the promising candidates for a universal memory. RRAM features a simple structure, small cell area, low switching voltage, and fast switching times. The resistive memory cell has a sandwiched structure with two metal electrodes above and below a metal oxide in the middle. To SET a cell, a positive voltage is applied across the device, increasing its conductance, i.e. switching to a low resistance state (LRS). To RESET a cell, a negative voltage is applied and the cell switches to a high resistance state (HRS). The cell retains the same resistance state even with no power supplied. Although the endurance is approaching 10^{10} cycles, it remains RRAM's primary challenge.

Conventionally, an RRAM cell is constructed of one transistor and one programmable resistive device (1T1R), as shown in Fig. 1. The transistor not only works as a switch for accessing the selected cell and isolating unselected ones, but also constrains the write current and limits cell disruption. However, in order to provide sufficient write current, the transistor needs to be large, which would dominate the cell area. An alternative approach is the crosspoint architecture [6], shown in Fig. 2(a). In a crosspoint array, RRAM cells are sandwiched between wordlines (WLs) and bitlines (BLs), which could achieve the ideal cell size of $4F^2$. Moreover, the resistive memory cells are fabricated in the back-end of the line (BEOL) process, which enables peripheral circuits to be hidden underneath the crosspoint array. Using a multi-layer structure [7] could further reduce the effective cell area, as shown in Fig. 2(b). However, the absence of access transistors in a crosspoint array complicates write and read operations.

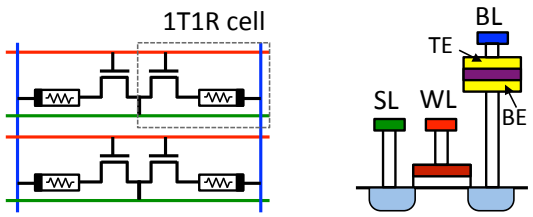


Fig. 1. 1T1R array and cell cross-sectional view.

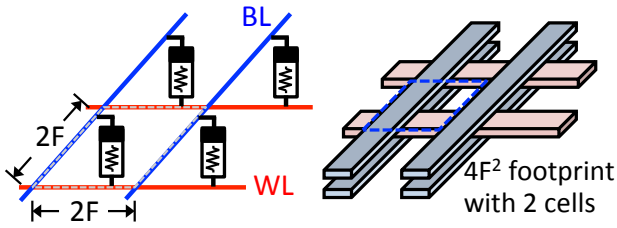


Fig. 2. (a) Crosspoint array, and (b) 2-layer structure.

The computing system memory hierarchy provides the illusion of a fast and large memory with high-speed, low-density caches and low-speed, high-density, large data storage. New, emerging nonvolatile memories, like RRAM, with sub-ns switching speed [8] have the potential to replace L2/L3 caches and eliminate the large standby leakage current.

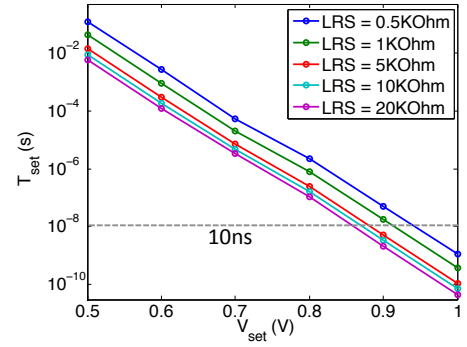
Section II describes the crosspoint architecture and its inherent issues. Section III proposes the differential 2R

crosspoint array. Section IV shows the circuit implementation of a 64KB crosspoint RRAM circuit including design techniques like array segmentation and the sense-before-write approach. Section V presents the simulation results. Section VI compares SRAM and RRAM as a cache in mobile applications. Conclusions are drawn in Section VII.

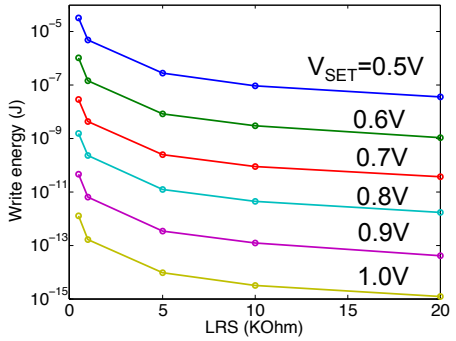
II. Crosspoint Array and Cell Analysis

A. RRAM Switching Behavior

The switching behavior of an RRAM cell depends on the write voltage (V_{SET}, V_{RESET}), the duration of write pulses (T_{SET}, T_{RESET}), and the high/low resistance values (R_H, R_L). Fig. 3(a) shows the tradeoff between the required time (T_{SET}) and voltage (V_{SET}) for programming a cell from the *HRS* to the *LRS* under different target R_L values. A higher R_L requires less time and energy to program and also suppresses the overall leakage current. However, to maintain a sufficient read margin, a smaller R_L is preferred so that the R_H vs. R_L ratio is larger. Fig. 3(b) shows the relationship between write energy and R_L under different V_{SET} *values*. Writing the cell with a higher voltage and a shorter pulse is more energy efficient. However, variations in the pulse duration widen the distribution of cell resistances.



(a)

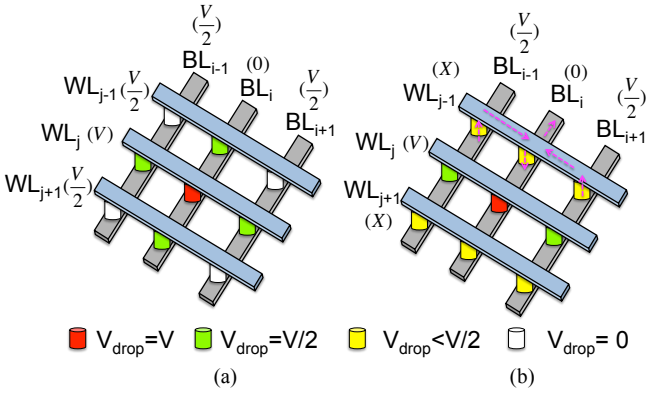


(b)

Fig. 3. (a) Write time and (b) write energy of a RRAM cell under different V_{SET} and R_L.

B. Leakage issues in crosspoint arrays

While a crosspoint array achieves high density by avoiding access transistors, it loses the ability to isolate unselected cells. To relax the requirements for minimizing write disturbance in crosspoint arrays, unselected WLs and BLs must be biased precisely. Fig. 4(a) shows the V/2 bias scheme, which limits the voltage disruption along the selected WL and BL to V/2. Another option is the floating wordline half-voltage bitline (FWHB) scheme shown in Fig. 4(b), which applies V/2 to the unselected BLs and floats the unselected WLs. In this case, the voltage drop across the cell (V_{drop}) is generally less than V/2, but it disturbs more cells. The write voltage should be large enough to successfully switch the cell but not so large as to cause a write disturbance. Undesired disruption voltages also induce leakage currents through unselected cells. The amount of leakage current is data-dependent, and the worst case occurs when all the unselected cells are in the LRS. Since the wire/switch resistance in an array is not negligible, variable IR drop amounts change the voltage applied across the cell, expand the cell variability distribution, and may even result in a write failure.



(a) (b)

Fig. 4. (a) V/2 bias scheme, and (b) FWHB scheme.

A common approach to detect the resistance state is current sensing, which mirrors the current flowing through the selected cell and compares it with a reference current (I_{REF}). However, the BL current (I_{BL}) in a crosspoint array includes both the selected cell current (I_{CELL}) and the total leakage current (I_{LEAK}). Fig. 5 illustrates the worst-case situation, when the selected cell is in a HRS and the other cells in the same array are all in the LRS. In this case, the read operation would fail when the BL current becomes larger than the reference current. Since the total leakage current depends on the number of cells, this situation constrains the array dimension. Also, the BL voltage fluctuation (ΔV_{BL}) and the leakage current are both data-dependent. Therefore, it is challenging to design a robust sensing circuit, under all cell variability distributions, data patterns, leakage currents, and PVT variations.

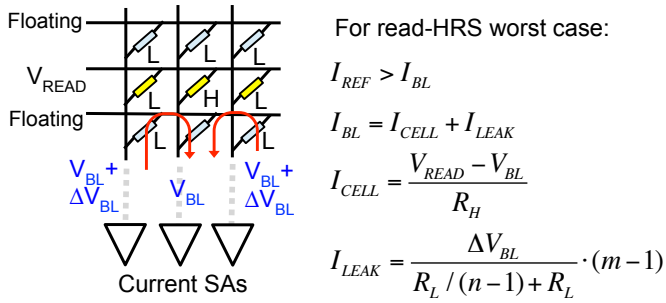


Fig. 5. Worst case of reading HRS in current sensing scheme. (m: BL length, n: WL length).

III. Differential 2R (D-2R) Cell and Crosspoint Array

We propose a differential 2R crosspoint structure, shown in Fig. 6(a), which can be read by using voltage sensing. The goal is to trade density for speed and robustness, thus to make it applicable for use in memory hierarchy. In this structure, two resistive devices with complementary resistance states are used to represent a 1-bit datum. To write a 1, SET R_T to a low resistance state and RESET R_B to a high resistance state; to write a 0, RESET R_T to a high resistance state and SET R_B to a low resistance state. The cell state can be readily determined by sensing the intermediate node X while applying V_{READ} to WL_T and ground to WL_B. The voltage on node X depends on the voltage divider formed by R_T and R_B. For evaluation purposes, BLs are connected to a StrongARM sense amplifier with a reference voltage of $V_{READ}/2$. Therefore, the read operation is immune to the leakage current flowing from neighboring BLs, which greatly increases the read margin without limiting the block size. The differential 2R cell contains both a HRS and a LRS, which solves the data dependency issue. Furthermore, the stacked resistors suppress leakage consumption during the read operation.

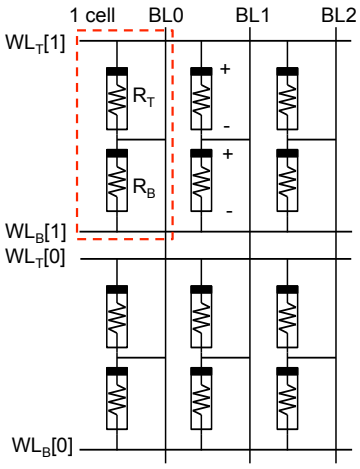
It is possible to design a 2R crosspoint array in a single layer of RRAM. However, thanks to the ability of stacking multiple RRAM layers, the differential 2R cell can be constructed between different metal layers with minimal area penalty. Since R_T and R_B have opposite electrodes connected to WL_T and WL_B, the same voltage can be applied to WL_T and WL_B to set one device and reset the other. The write operation is illustrated in Fig. 6(b). In the write-1 operation, both WL_T and WL_B are connected to a write voltage, V_{write}, and the BL is connected to ground. A positive V_{write} drops across R_T, which sets R_T to the LRS. In the meantime, a negative V_{write} drops across R_B, which resets R_B to the HRS. In contrast, to write a zero, the BL is connected to V_{write}, and WL_T and WL_B are connected to ground.

The forming operation in the initialization step is required to construct the conductive filament in each resistive device after fabrication. A forming operation is similar to a set operation with a higher voltage and a longer duration. Two sequential phases are applied to initialize R_T and R_B separately. In the first phase, the selected WL_T is connected to V_{form} while the selected BL and WL_B are held at ground. In the second phase, WL_T and the BL are connected to V_{form} and WL_B is connected to ground. In the two phases, R_T and R_B are applied to V_{form} and switched to the LRS respectively.

IV. Circuit Implementation

A. Array Segmentation

There are twice as many cells in the D-2R array as in the conventional crosspoint array. During operation, half of the cells are in the HRS and half of them are in the LRS. Therefore, the leakage current would be 8% larger than in the worst case of a conventional crosspoint array. However, the leakage current is a constant value in the D-2R scheme, and the data-dependent variable IR drop issue does not exist.



	Form-R_T	Form-R_B
R_T	FORM	X
R_B	X	FORM
WL_T	V_{FORM}	V_{FORM}
WL_B	0	0
BL	0	V_{FORM}

	Write-1	Write-0
R_T	SET	RESET
R_B	RESET	SET
$WL_{T/B}$	V_{WRITE}	0
BL	0	V_{WRITE}

Read operation:
$WL_T = V_{READ}$, $WL_B = 0$
$V_{BL}=V_{READ}\times R_B/(R_T+R_B)$

Fig. 6. (a) Differential 2R crosspoint array and (b) table of operating conditions in form/write/read mode.

The write current (I_{WRITE}) in the D-2R scheme with V/2 biasing includes the cell current ($I_{CELL}=V/R_L$) and the leakage current ($I_{LEAK} \approx$ (n-1)$\times V/R_L$). The energy efficiency (I_{CELL}/I_{WRITE}) decreases with increasing array dimensions. Array segmentation, similar to the divided wordline technique [9] employed in SRAM for reducing WL loading, disturbance, and power consumption, is used here to reduce the number of activated cells and mitigate the write leakage current. To keep the write current under 100μA, 4-cell wide WLs are required. Instead of building a 4x4 array with its own peripheral circuit, a large array is constructed by segmenting one WL into local WLs (LWLs). Only one LWL is active at a time to reduce the write leakage current. Switches are inserted every four columns to connect the global WL (GWL) and LWLs. Fig. 7 shows a cross-sectional view of the D-2R array with array segmentation. Although placing transistors under the array minimizes their overhead, additional area is consumed for routing transistors to the GWL and LWL metal layers. There is a tradeoff between area penalty and leakage current. For a LWL of 4 cells wide, the area would be twice the size of that without array segmentation. Compared to a 140 F^2 SRAM bit cell, an RRAM cell in a crosspoint array of $4F^2$ cell area is 35x smaller. However, the area penalty due to array segmentation increases the equivalent bit cell area to $10F^2$, which is still much smaller than an SRAM bit cell and a 1T1R RRAM cell.

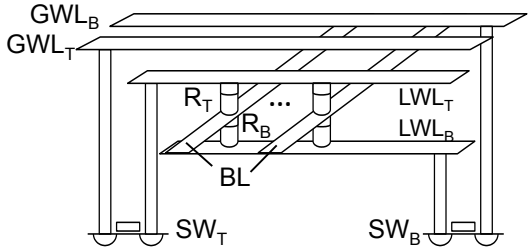


Fig. 7. Cross-sectional view of the differential 2R array with array segmentation.

B. Sense-Before-Write

The resistive value of the memory cell varies with the voltage and period of the write pulse. Repeated SET pulses

applied to the same cell reduce its resistance value until it hits the lowest resistance level. Doing this would result in very high current consumption and a wide cell resistance distribution. To prevent the over-SET situation, the sense-before-write approach is applied [10]. At the beginning of the write cycle, a read operation is first conducted and the output is fed back to a control circuit to determine whether to write or not. The cell would not be written again unless there is a need to flip the state. By using the sense-before-write approach, the cell resistance distribution is narrower and the leakage current is suppressed by keeping each LRS at a higher resistance value. Moreover, avoiding unnecessary cell access elongates the endurance.

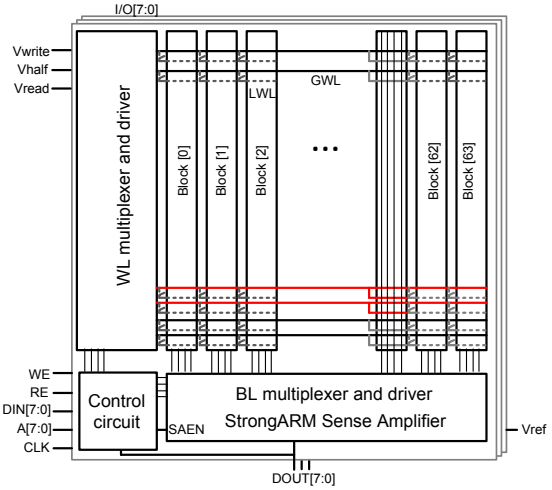


Fig. 8. Block diagram of a 64KB crosspoint RRAM circuit.

The block diagram of a 64KB D-2R crosspoint RRAM macro that contains 8 blocks is shown in Fig. 8. In SRAM, WLs drive the gates of access transistors. In a crosspoint array, however, WLs are connected to V_{write}, V_{read}, or ground, depending on data input values and operational modes. Therefore, 8 bits of data need separate WL/BL drivers to provide the correct voltage to program the cell. V_{write}, V_{read} and the unselected BL voltage (V_{half}) are provided by the voltage generator, which is not shown in the block diagram.

The control circuit generates all the input control signals, such as write enable (WE), read enable (RE), input data (DIN), addresses (A) and output data (DOUT), to determine the operational mode and corresponding control signals to read/write circuits. WL/BL multiplexers and drivers deliver different voltage levels (V_{form}, V_{write}, V_{half}, V_{read} and ground) to WLs and BLs according to the control signals.

The read voltage (V_{read}) is set to a low value of 0.3V to prevent disturbance. Thus, the StrongARM sense amplifiers with PMOS input transistors are used to sense the inputs with low common mode. It compares the BL voltage to the reference voltage (V_{ref}) and outputs the result. The voltage-sensing scheme in D-2R crosspoint array is less susceptible to cell distribution and data pattern variability than the conventional current-sensing scheme in a 1R crosspoint array.

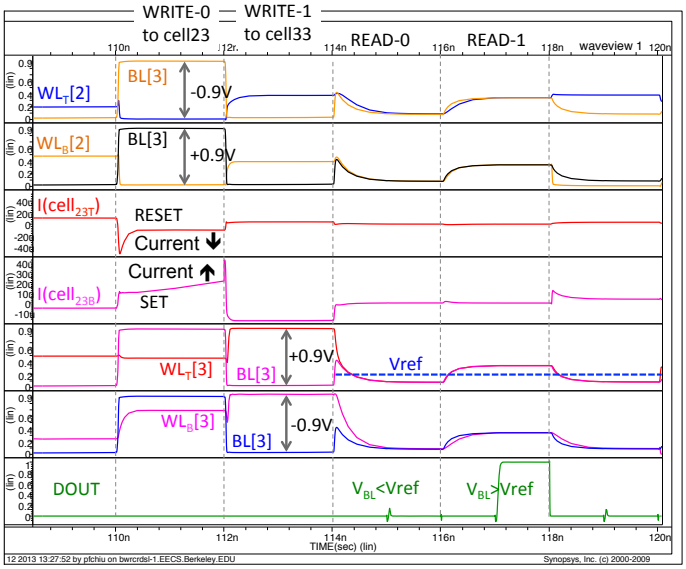


Fig. 9. Waveform of read and write operation in D-2R crosspoint array.

V. Simulation Results

Simulation of one block and its peripheral circuits is conducted using Eldo with a 28/32nm predictive technology model (PTM) and a Verilog-A RRAM model. The RRAM model illustrates the physical behavior of SET/RESET processes to fit the measurement results [11, 12]. Fig. 9 shows the simulation waveform. In the highlighted period, WL_T[2] and WL_B[2] are connected to ground and BL[3] is connected to V_{write} to SET $cell_{23B}$ and RESET $cell_{23T}$. The unselected WLs are kept at $V_{write}/2$ to prevent disturbance. The switching behavior of $cell_{23T}$ and $cell_{23B}$ is confirmed by noting the increase in current of $cell_{23B}$ (SET operation) and the decrease in current of $cell_{23T}$ (RESET operation).

During a read operation, V_{read} is applied to the selected WL_T and WL_B is connected to ground. Thus, the BL voltage is proportional to the resistance ratio of R_T and R_B. The sense amplifier compares this BL voltage to V_{ref} to determine DOUT. The sense enable (SAEN) signal is triggered after the voltage difference is fully developed.

Table I shows the parameters used for simulating the D-2R crosspoint RRAM circuit. The average current during a write cycle in each block is 140 µA, and the average current during a read cycle in each block is 17 µA. The switches are designed for a maximum voltage drop of 50mV during read/write.

Table I. Parameters in D-2R circuit simulation.

Clock Frequency	500 MHz
Density	64KB
Power supply	1.0 V
Write voltage (V_{write})	0.95 V
Read voltage (V_{read})	0.3 V
Reference voltage (V_{ref})	0.2 V
R_H/R_L	90KΩ/8KΩ
Write current (one block)	140 µA
Read current (one block)	17 µA
Standby current	~ 0 mA

Table II. Comparisons between various memory technologies for cache usage.

Clock Frequency	SRAM	eDRAM	STT-RAM	1R crosspoint	D-2R
Cell size (F^2)	100-200	20-50	6-50	4	10
Write energy	Low	Low	High	High	High
Read/Write speed	High	Medium	Low	Low	Low
Standby leakage	High	Low	None	None	None
Endurance	High	High	High	Medium	Medium
Retention time	-	<100us	Nonvolatile	Nonvolatile	Nonvolatile
Features	High speed	Small cell size	High endurance	Small cell size	Higher read margin
Challenge	Leakage	Refresh	Yield	Sensing error	Power consumption

VI. Differential RRAM in Memory Hierarchy

The process variation in advanced technologies prevents the scaling of SRAM bit cells. The area overhead and leakage energy consumption are significant for on-chip last-level cache. To reduce miss penalty by increasing memory capacity, eDRAM provides an option of high-density cache memory. The bit cell area is 20-50 F^2 [13], which is about 4x smaller than an SRAM bitcell. However, an extra process to add the capacitors and the need of refresh cycles increase cost and energy. RRAM is another approach to reach high density. Ideally, the bit cell size is $4F^2$ in crosspoint array and $10F^2$ in D-2R crosspoint structure. Moreover, the non-volatility eliminates the leakage current of high-capacity last-level cache. Therefore, nonvolatile cache is attractive for long-standby battery-driven consumer devices. Aside from non-volatility, the potential of stacked layers enables even larger memory capacity. A comparison table of various memory technologies for cache usage is provided in Table II.

RRAM endurance of 10^{10} is below the 10^{16} requirement for the conventional L3 cache. However, it meets the needs of a context-switching memory in mobile systems [14]. Contexts of idle applications, which reside in storage to mitigate power consumption, take a long time to recall while users switch over different applications. It requires orders of magnitude more reads than writes and is of growing importance in mobile computing. Parallel read is feasible to further increase the read throughput, which greatly improves the performance for context switch purpose.

The endurance can be improved by system or circuit approaches. In addition to the sense-before-write scheme, wear-leveling spreads the write operations evenly across the memory and the built-in test circuit monitors the worn cell status.

VII. Conclusion

In this work, we propose a voltage-sensing differential 2R crosspoint structure. It enhances the read margin and solves the sensing error due to leakage in a current-sensing scheme. Also, having the same number of HRS and LRS cells prevents data pattern problems and avoids variable IR drop. To avoid disturbance and limit the leakage current during a write operation, an array segmentation scheme with WL length of 4 cells wide is adopted. This constrains the write current to below 200μA. The sense-before-write approach prevents cells from having variable LRS resistance values and constrains the cell variability distribution.

A 64KB differential 2R crosspoint RRAM memory can operate at 500 MHz with an average write current of 140 μA and an average read current of 16.6 μA. The sense-before-write scheme requires two cycles to complete a write operation. Also, the array segmentation scheme suffers 2x area penalty but effectively reduces the leakage current.

An envisioned application as a context memory presents an attractive application for the D-2R crosspoint RRAM. The equivalent cell size is 10 F^2, much smaller than an SRAM bit cell. Elimination of the standby current outweighs the higher write energy.

Acknowledgement

This work was funded in part by DARPA PERFECT program, contract HR0011-12-2-0016. The authors would like to thank Olivier Thomas and Natalija Jovanovic from CEA-Leti, and Jean-Michel Portal and Marc Bocquet from IM2NP.

References

[1] Y. Li, et al., "128Gb 3b/cell NAND flash memory in 19nm technology with 18MB/s write rate and 400Mb/s toggle mode," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 436-437.

[2] T. Takashima, et al., "A 100MHz ladder FeRAM design with capacitance-coupled-bitline (CCB) Cell," *IEEE Journal of Solid-State Circuits*, Vol. 46, No. 3, March 2011.

[3] D. C. Ralph and M. D. Stiles, "Spin transfer torques," *Journal of Magnetism and Magnetic Materials*, vol. 320, issue 7, pp. 1190-1216, April 2008.

[4] R. E. Simpson, et al., "Toward the ultimate limit of phase change in $Ge_2Sb_2Te_5$," *Nano Letter*, pp. 414-419, 2010.

[5] R. S. Williams, "How we found the missing memristor," *IEEE Spectrum*, vol. 45, no. 12, pp. 28-35, 2008.

[6] E. Ou and S. S. Wong, "Array architecture for a nonvolatile 3-dimensional cross-point resistance-change memory," *IEEE J. Solid-State Circuits*, vol. 46, no. 9, pp. 2158-2170, Sep. 2011.

[7] A. Kawahara, et al., "An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput," *IEEE Journal of Solid-State Circuits*, Vol. 48, No. 1, January 2013.

[8] H. Y. Lee et al., "Evidence and solution of over-RESET problem for HfO_x based resistive memory with sub-ns switching speed and high endurance," in *Int. Electron Devices Meeting (IEDM) Tech. Dig. Papers*, Dec. 2010, pp. 460–463.

[9] M. Yoshimoto, et al., "A Divided word-line structure in the static SRAM and its application to a 64K full CMOS RAM" *IEEE Journal of Solid-State Circuits*, Vol. 18, No. 5, pp. 479–485, Oct. 1983.

[10] J. Ahn and K. Choi, "Lower-bits cache for low power STT-RAM Caches," in *IEEE Int. Symp. on Circuits and Systems (ISCAS)*, May 2013, pp. 480–483.

[11] C. Cagli, et al., "Experimental and theoretical study of electrode effects in HfO_2 based RRAM," in *Int. Electron Devices Meeting (IEDM) Tech. Dig. Papers*, Dec. 2011, pp. 658–661.

[12] M. Bocquet, "Robust compact model for bipolar oxide-based resistive switching memories," in *IEEE Transactions on Electron Devices*, Vol. 61, No. 3, pp. 674–681, March 2014.

[13] F. Hamzaoglu, et al., "A 1Gb 2GHz embedded DRAM in 22nm tri-gate CMOS Technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 230–231.

[14] H. Kim, et al., "Revisiting storage for smartphones," in *Proceedings of the 10th USENIX File and Storage Technologies (FAST)*, 2012, pp. 17–31.