A Differential 2R Crosspoint RRAM Array With Zero Standby Current

Pi-Feng Chiu, *Student Member, IEEE*, and Borivoje Nikolić, *Senior Member, IEEE*

***Abstract*—Memory power consumption dominates mobile system energy budgets in scaled technologies. Fast nonvolatile memory devices (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 brief, we propose a differential 2R crosspoint structure with array segmentation and sense-before-write techniques. A 64-KB RRAM device is constructed and simulated in a 28/32-nm CMOS predictive technology model and a Verilog-A RRAM model. This design offers an opportunity to use RRAM as a cache for increasing energy efficiency in mobile computing.**

***Index Terms*—Cache, crosspoint, memristor, nonvolatile memory (NVM), resistive random access memory (RRAM), 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 memory devices will dominate energy consumption in standby mode. Nonvolatile memory devices (NVMs) can be powered down completely, eliminating the leakage current. Flash memory [1], the most popular NVM, 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 NVM that can be used as a universal memory device, replacing both Flash memory and SRAM.

New memory technologies include ferroelectric random access memory (FeRAM) [2], spin–transfer torque random access memory (STT-RAM) [3], phase-change random access memory (PRAM) [4], and resistive random access 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 a high switching speed, and it 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 device. 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.

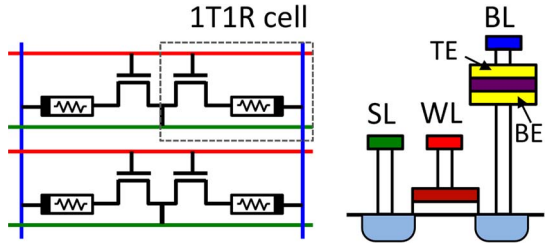


Fig. 1. 1T1R array and cell cross-sectional view.

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], as 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 multilayer 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.

The computing system memory hierarchy provides the illusion of a fast and large memory device with high-speed low-density caches and low-speed, high-density, and large data storage. New and emerging NVMs, such as RRAM, with subnanosecond switching speed [8] have the potential to

Manuscript received July 14, 2014; revised September 30, 2014; accepted December 16, 2014. Date of publication December 23, 2014; date of current version April 23, 2015. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) through the Power Efficiency Revolution For Embedded Computing Technologies (PERFECT) program under Contract HR0011-12-2-0016. This brief was recommended by Associate Editor P. Li.

P.-F. Chiu is with Berkeley Wireless Research Center, University of California at Berkeley, Berkeley, CA 94704-1302 USA (e-mail: pfchiu@eecs.berkeley.edu).

B. Nikolić is with the Berkeley Wireless Research Center, University of California at Berkeley, Berkeley, CA 94704-1302 USA (e-mail: bora@eecs.berkeley.edu).

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

Digital Object Identifier 10.1109/TCSII.2014.2385431

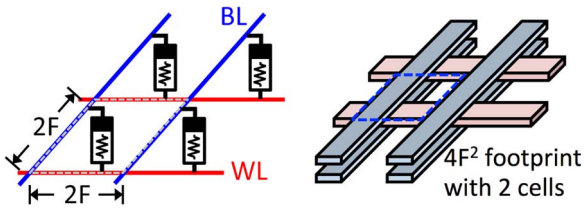


Fig. 2. (a) Crosspoint array. (b) Two-layer structure.

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 (D-2R) crosspoint array. Section IV shows the circuit implementation of a 64-KB crosspoint RRAM circuit, including design techniques such as 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. The conclusion is 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 $(V_{\rm SET}, V_{\rm RESET})$, the duration of write pulses $(T_{\rm SET}, T_{\rm RESET})$, and the high/low resistance values (R_H, R_L). Fig. 3(a) shows the tradeoff between the required time $T_{\rm SET}$ and voltage $V_{\rm SET}$ 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 versus R_L ratio is larger. Fig. 3(b) shows the relationship between write energy and R_L under different $V_{\rm SET}$ *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.

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 WL half-voltage BL (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 $V_{\rm drop}$ 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 very 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 common approach to detect the resistance state is current sensing, which mirrors the current flowing through the selected

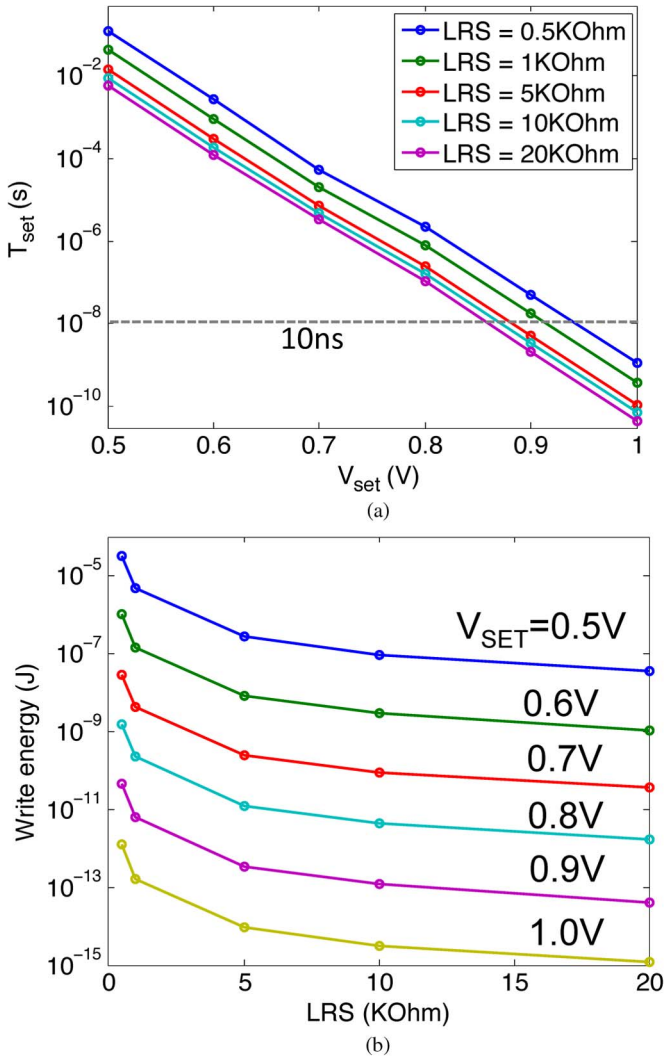


Fig. 3. (a) Write time and (b) write energy of an RRAM cell under different $V_{\rm SET}$ and R_L values.

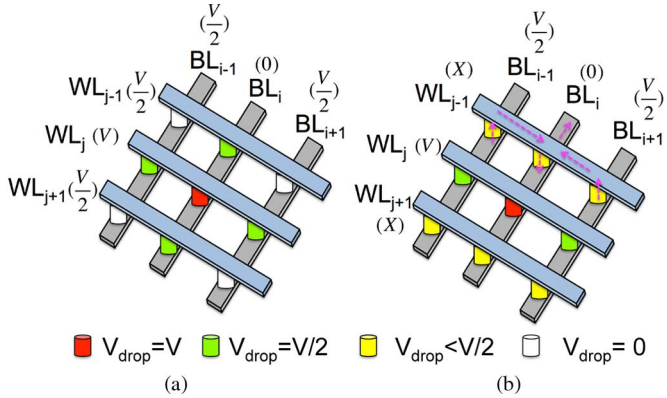
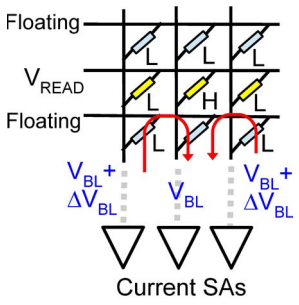


Fig. 4. (a) V/2 bias scheme. (b) FWHB scheme.

cell and compares it with a reference current $I_{\rm REF}$. However, the BL current $I_{\rm BL}$ in a crosspoint array includes both the selected cell current $I_{\rm CELL}$ and the total leakage current $I_{\rm LEAK}$. Fig. 5 illustrates the worst case situation when the selected cell is in an 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



For read-HRS worst case:

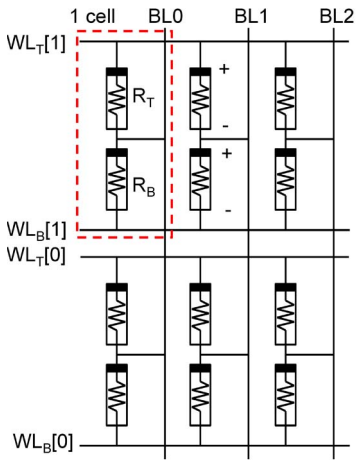
$$I_{REF} > I_{BL}$$

$$I_{BL} = I_{CELL} + I_{LEAK}$$

$$I_{CELL} = \frac{V_{READ} - V_{BL}}{R_H}$$

$$I_{LEAK} = \frac{\Delta V_{BL}}{R_L/(n-1) + R_L} \cdot (m-1)$$

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



	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) D-2R crosspoint array. (b) Table of operating conditions in form/write/read mode.

situation constrains the array dimension. In addition, 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 process–voltage–temperature variations.

III. D-2R Cell and Crosspoint Array

We propose a D-2R crosspoint structure, as 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 an LRS and RESET R_B to an HRS; to write a 0, RESET R_T to an HRS and SET R_B to an LRS. 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 $V_{\mathrm{READ}}/2$. 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 D-2R cell contains both an HRS and an LRS, which solves the data dependence 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, due to the ability of stacking multiple RRAM layers, the D-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.

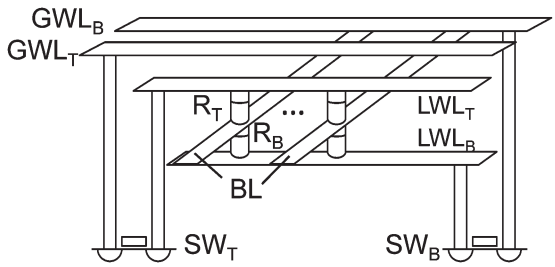


Fig. 7. Cross-sectional view of the D-2R array with array segmentation.

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 period. 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.

The write current I_{WRITE} in the D-2R scheme with V/2 biasing includes the cell current ($I_{\mathrm{CELL}} = V/R_L$) and the leakage current ($I_{\mathrm{LEAK}} \approx (n-1) \times V/R_L$). Energy efficiency ($I_{\mathrm{CELL}}/I_{\mathrm{WRITE}}$) decreases with increasing array dimensions. Array segmentation, similar to the divided WL technique [9] employed in SRAM to reduce 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 100 μA, four-cell-wide WLs are required. Instead of building a 4×4 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

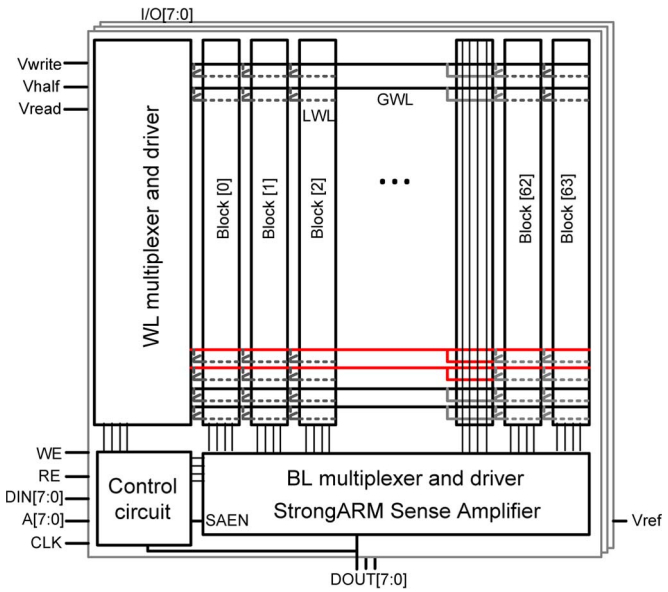


Fig. 8. Block diagram of a 64-KB crosspoint RRAM circuit.

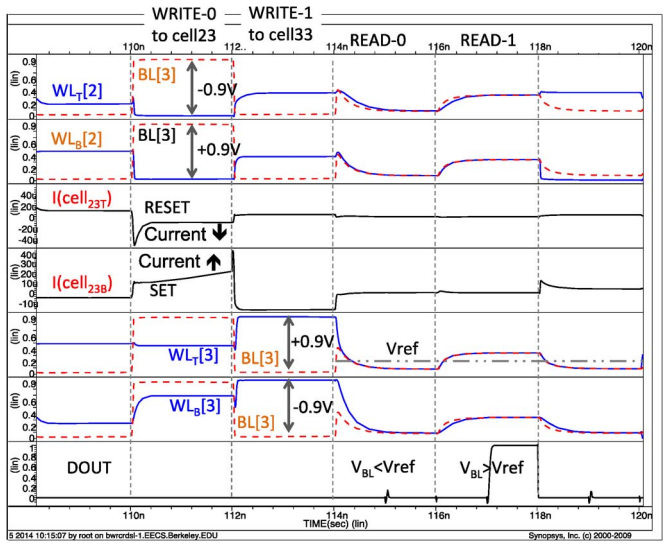


Fig. 9. Waveform of read and write operations in the D-2R crosspoint array.

the GWL and LWL metal layers. There is a tradeoff between area penalty and leakage current. For an LWL of four cells wide, the area would be twice the size of that without array segmentation. Compared with a 140-F^2 SRAM bit cell, an RRAM cell in a crosspoint array of 4-F^2 cell area is 35× smaller. However, the area penalty due to array segmentation increases the equivalent bit cell area to 10 F^2, which is still much smaller than that of an SRAM bit cell and a 1T1R RRAM cell.

B. Sense-Before-Write Approach

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.

The block diagram of a 64-KB D-2R crosspoint RRAM macro that contains eight blocks is shown in Fig. 8. In SRAM, WLs drive the gates of access transistors. In a crosspoint array, however, WLs are connected to $V_{\rm write}$, $V_{\rm read}$, 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. $V_{\rm write}$, $V_{\rm read}$, and the unselected BL voltage $V_{\rm half}$ 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 (i.e., $V_{\rm form}$, $V_{\rm write}$, $V_{\rm half}$, $V_{\rm read}$, and ground) to WLs and BLs according to the control signals.

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

Read voltage $V_{\rm read}$ is set to a low value of 0.3 V 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 $V_{\rm ref}$ and outputs the result. The voltage-sensing scheme in the 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.

V. Simulation Results

Simulation of one block and its peripheral circuits is conducted using Eldo with a 28/32-nm 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 $V_{\rm write}$ to SET $\rm cell_{23B}$ and RESET $\rm cell_{23T}$. The unselected WLs are kept at $V_{\rm write}/2$ to prevent disturbance. The switching behavior of $\rm cell_{23T}$ and $\rm cell_{23B}$ is confirmed by noting the increase in current of $\rm cell_{23B}$ (SET operation) and the decrease in current of $\rm cell_{23T}$ (RESET operation).

During a read operation, $V_{\rm read}$ 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 $V_{\rm ref}$ to determine DOUT.

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

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 50 mV during read/write.

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 4$\times$ smaller than an SRAM bit cell. 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 4 F^2 in the crosspoint array and 10 F^2 in the D-2R crosspoint structure. Moreover, nonvolatility eliminates the leakage current of high-capacity last-level cache. Therefore, nonvolatile cache is attractive for long-standby battery-driven consumer devices. Aside from nonvolatility, the potential of stacked layers enables even larger memory capacity. A comparison 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 brief, we have proposed a voltage-sensing D-2R crosspoint structure. It enhances the read margin and solves the sensing error due to leakage in a current-sensing scheme. In addition, 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 four cells wide is adopted. This constrains the write current to below 200 μA. The sense-before-write approach prevents cells from having variable LRS resistance values and constrains the cell variability distribution.

A 64-KB D-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. In addition, the array segmentation scheme suffers 2$\times$ area penalty but effectively reduces the leakage current.

An envisioned application as a context memory device 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.

ACKNOWLEDGMENT

The authors would like to thank O. Thomas and N. Jovanovic from CEA-Leti and J.-M. Portal and M. Bocquet from IM2NP.

REFERENCES

[1] Y. Li *et al.*, "128 Gb 3b/cell NAND Flash memory in 19 nm technology with 18 MB/s write rate and 400 Mb/s toggle mode," in *Proc. IEEE ISSCC Dig. Tech. Papers*, Feb. 2012, pp. 436–437.

[2] T. Takashima, Y. Nagadomi, and O. Tohru, "A 100 MHz ladder FeRAM design with capacitance-coupled-bitline (CCB) cell," *IEEE J. Solid-State Circuits*, vol. 46, no. 3, pp. 681–689, Mar. 2011.

[3] D. C. Ralph and M. D. Stiles, "Spin transfer torques," *J. Magn. Magnetic Mater.*, vol. 320, no. 7, pp. 1190–1216, Apr. 2008.

[4] R. E. Simpson *et al.*, "Toward the ultimate limit of phase change in $Ge_2Sb_2Te_5$," *Nano Lett.*, vol. 10, no. 2, pp. 414–419, Feb. 2010.

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

[6] E. Ou and S. S. Wong, "Array architecture for a nonvolatile three-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 8 Mb multi-layered cross-point ReRAM macro with 443 MB/s write throughput," *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 178–185, Jan. 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 *Proc. IEEE IEDM*, Dec. 2010, pp. 460–463.

[9] M. Yoshimoto *et al.*, "A divided word-line structure in the static SRAM and its application to a 64 K full CMOS RAM" *IEEE J. 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 *Proc. IEEE ISCAS*, May 2013, pp. 480–483.

[11] C. Cagli *et al.*, "Experimental and theoretical study of electrode effects in HfO_2 based RRAM," in *Proc. IEEE IEDM*, Dec. 2011, pp. 658–661.

[12] M. Bocquet, "Robust compact model for bipolar oxide-based resistive switching memories," *IEEE Trans. Electron Devices*, vol. 61, no. 3, pp. 674–681, Mar. 2014.

[13] F. Hamzaoglu *et al.*, "A 1 Gb 2 GHz embedded DRAM in 22 nm tri-gate CMOS technology," in *Proc. IEEE ISSCC Dig. Tech. Papers*, Feb. 2014, pp. 230–231.

[14] H. Kim *et al.*, "Revisiting storage for smartphones," in *Proc. 10th USENIX FAST*, 2012, pp. 17–31.