# Energy Efficiency Enhancement in Mobile Processor Memory Design Using Emerging Nonvolatile Memory

Pi-Feng Chiu, Pengpeng Lu, and Zeying Xin
{pfchiu, penpenglu, xinzeying}@berkeley.edu

***Abstract*— The leakage power consumption of memories is becoming dominant in a system as technology scaling down. Fast nonvolatile memories (NVMs) offer a tremendous opportunity to eliminate memory leakage current during standby mode. Resistive random access memory (RRAM) is considered to be one of the most promising emerging NVMs due to high speed, small area, and low power consumption. Crosspoint structure makes RRAM array to achieve high density. However, this put big challenges to the peripheral circuit design. In this paper, several design techniques of crosspoint structure are explored. Also, energy analysis of RRAM is conducted and compared with SRAM. Finally, we will discuss the possibility of utilizing RRAM as cache to increase energy efficiency in mobile applications.**

***Keywords*— cache, crosspoint, energy analysis, fast power-on, mobile system, NVM, RRAM**

## I. Introduction

Memories have been the largest portion in integrated circuit of consumer electronics in terms or area and energy consumption. As the trend of scaling technology goes, the leakage current issue in memories becomes more and more severe. There have been a lot of design techniques to tackle this problem, such as dynamic voltage scaling (DVS). However, to thoroughly eliminate idling current, NVMs are the best choice, which can be completely shut down without worrying about the loss of data. Moreover, utilizing NVM as cache allows us to achieve instant power-on procedure without the need to transfer data to main memory. Flash memory is the most popular NVM in the market due to the small cell size. However, it could never replace SRAM as a cache because of the slow write speed. Flash memory also has other inherent issues. First of all, the program and erase mechanism requires a high voltage, typically greater than 5V, which degrades reliability and endurance. Most commercially available flash products are guaranteed to withstand around 100,000 P/E cycles before the wear begins to deteriorate the integrity of the storage [1]. Power-hungry and large area circuit like charge pump is needed to generate the high voltage. Also, flash memory is facing the physical limitation that it can hardly be scaled down further more. Below the minimal geometric limitation, the yield and reliability could decrease dramatically [2, 3]. Those limitations prevent flash memory from monopolizing NVM market in the future development. Therefore, a new nonvolatile memory needs to be developed to replace flash memory with comparable yield and read/write speed.

In recent year, there have been several emerging nonvolatile memories developing. Ferroelectric RAM (FeRAM) is a random access memory similar in construction to DRAM, but uses a ferroelectric layer instead of a dielectric layer to achieve non-volatility. FeRAM has the advantages of faster write speed and a better endurance than flash memory. However, it hasn't come to mass-production because of the lower storage density, which means higher cost. Magnetoresistive random access memory (MRAM) stores data by using a fixed layer and a free layer, one is a permanent magnet set to a particular polarity, the other's field can be changed to match that of an external field to store memory, respectively. This approach requires a fairly substantial current to generate the field, however, which makes it less interesting for low-power uses, one of MRAM's primary disadvantages. Spin-transfer torque random access memory, or STT-RAM, has the advantages of lower power consumption and better scalability over conventional MRAM. However, the amount of current needed to reorient the magnetization is still too high for most commercial applications [4]. Phase-change memory (PRAM)'s most appealing point is the switching time and inherent scalability [5]. The temperature sensitivity is perhaps its most notable drawback, one that may require changes in the production process of manufacturers incorporating the technology. Conductive-bridging RAM (CBRAM) is based on the physical relocation of ions within a solid electrolyte. In contrast to flash memory, CBRAM writes with a relatively low power and at high speed.

Among all these emerging nonvolatile memories, one of the most promising candidates is the resistive random access memory (RRAM). The basic idea is that a dielectric, which is normally insulating, can be made to conduct through a filament or conduction path formed after application of a sufficiently high voltage. Once the filament is formed, it may be RESET (broken, resulting in high resistance) or SET (re-formed, resulting in lower resistance) by an appropriately applied voltage. The features including low voltage, small cell area, and fast write time make RRAM the hottest topic to investigate. To further increase the memory density, crosspoint structure is employed. Since the cells are fabricated in back end of line (BEOL), all the peripheral circuits can be hidden under the cell array. Furthermore, multilayer array can be realized to maximize array efficiency. However, the absence of controlling selector in crosspoint array means more challenges on the peripheral circuit design.

In this paper, we will review several schemes in crosspoint array. A design flow will be built to choose optimal operation conditions and block size. The main objective is to construct a functional RRAM circuit with crosspoint structure. Some simulation results will be shown to investigate the design tradeoffs. Also, we will conduct the energy analysis to RRAM

and compare it with SRAM. Finally, we will discuss the possibility of utilizing RRAM as cache to increase energy efficiency in mobile applications.

## II. Array Structure

There are two possible memory structures for a RRAM array: 1Selector1Resistor (1S1R) and crosspoint structure. The selector can be realized by MOSFETs, BJTs or diodes. In the most common 1T1R structure, the memory cell consists of a resistive memory element and a MOSFET as a selector. As the size of a MOSFET access device is typically much larger than the size of a RRAM cell, the total area of memory array is primarily dominated by MOSFETs rather than by RRAM cells. Also, in order to provide enough drive current, larger than minimum-sized transistor should be used for write operations. Hence, RRAM's area advantage goes down significantly because of the access devices.

Instead of the big MOSFET, we can use diode as the cell select element. The resistance of RRAM significantly increases as the voltage applied on it decreases, which indicates effective cut off of the leakage current from the unselected cells in the sneak paths. Therefore, the area-efficient crosspoint RRAM is enabled. Fig. 1 shows the array structure of 1T1R, crosspoint, and 4-layer crosspoint [6]. In crosspoint array, RRAM cells are sandwiched between wordlines (WLs) and bitlines (BLs), and the actual cell area can be reduced by around 1/4. Moreover, since a crosspoint array permits a multilayered structure, its effective cell area is further reduced, for example, to 1/16 in case of 4 layers. Although avoiding access transistor is beneficial from cell area standpoint, it introduces other complexities. Some papers have discussed about the issues in crosspoint array, which can be separated into two categories, write operation and read operation.

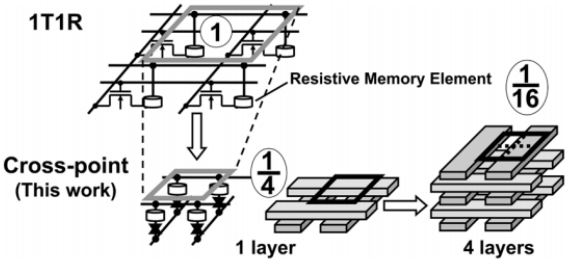



Fig. 1. Array structure of 1T1R, crosspoint and 4-layer crosspoint. [6]

### A. Write operation

Write reliability is a serious concern in crosspoint arrays. In an ideal condition, the resistance of wires and the sneak currents in unselected cells are negligible. In such a scenario, the write voltage $V_{WL}$-$V_{BL}$ is fully applied across the specified cell. However, in reality, both wire resistance and sneak current are non-trivial. Hence, the voltage applied across a cross-point varies based on the location of the cell as well as the data pattern stored in all of the RRAM cells in the array. A write is considered reliable if it modifies the content of the selected cells to the new value without disturbing other unselected cells. Correspondingly, there are two potential problems with writes: *write failure*, an unsuccessful write on selected cells, and *write disturbance*, and undesirable write to unselected cells.

Fig. 2 shows a common accessing method, V/2 biasing scheme [7]. When cell is selected during a write operation, other memory cells connected to the selected BL and WL are subject to half of the write voltage and defined as half-select cells.

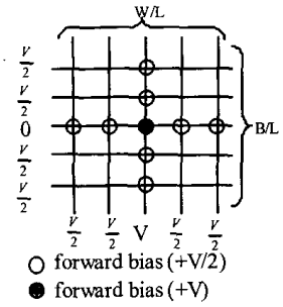



Fig. 2. V/2 biasing scheme. Solid circle cells are under forward bias of V, hollow circles are under bias of V/2, other cells are free from voltage bias. [7]

When multiple cells are selected in a single WL, it is impossible to finish the write operations (both SET and RESET) in one step, since no reasonable voltage on the unselected WL can be applied: if the voltage on the unselected WL is smaller than that on the selected WL, the voltage applied on the cell between the unselected WL and the selected BL during RESET ($V_{BL}$-$V_{WL}$=$V_{RESET}$) will be higher than the RESET voltage, thus the unselected cells will be undesirably RESET; if the voltage on the unselected WL is larger than that on the selected WL, the cell between the unselected WL and the selected BL during SET will be also be SET undesirably. Therefore, SET and RESET operations cannot be performed simultaneously in the same crosspoint structure. We can directly separate SET and RESET operations, or we can SET all the columns first and erase the unwanted ones later [8]. Fig. 3 and Fig. 4 illustrate the idea of sequential write method.

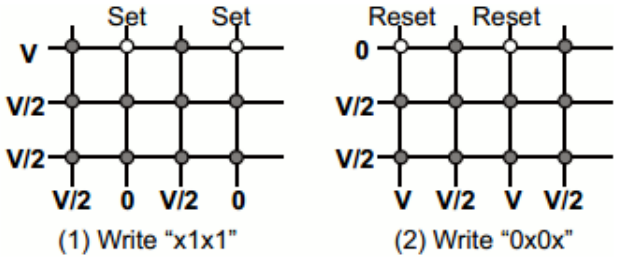



Fig. 3. Sequential write method: SET-before-RESET. [8]

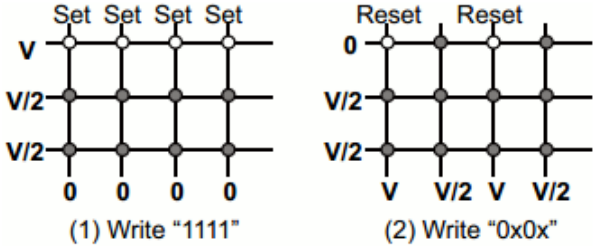



Fig. 4. Sequential write method: ERASE-before-RESET. [8]

### B. Read operation

To apply a read voltage on the selected cell, other half-select cells generate sneak current, and may lead to a read failure, especially when the selected cell is in high resistance state ($R_H$) and the half-select cells are in low resistance state ($R_L$). In order to alleviate read disturbance, parallel read can be employed to read all the cells in the same row to eliminate half-select issue. The other circuit-level solution is a two-step sampling method, which could isolate the noise current from the parasitic half select cells. We first read the background current of the half-select cells and latch it. In the next step, the total current of the background current plus the current through the selected memristor cell is read. Finally, after removing of the background current, the state of the selected memory cell is identified by the sensing scheme [8].

## III. Analysis of RRAM Cell and Array

Before starting to build a crosspoint circuit, some cell parameters are required to determine write/read voltages ($V_{SET}$, $V_{RESET}$, $V_{READ}$), period of write pulses ($T_{SET}$, $T_{RESET}$), and high/low resistance values ($R_H$, $R_L$). The information can be extracted from the physical model characterizing RRAM behavior. Note that cell distribution is also an important factor, which would be discussed later. According to the RRAM model, we can analyze the cell characteristics to get the optimal block size during write and read operation. Fig. 6 shows the required time ($T_{SET}$) to set the cell from $R_H$ to $R_L$ under different targeted $R_L$ value, which reveals the tradeoff between write voltage and write time. Fig. 7 plots the relationship between write energy and $R_L$ value under different $V_{SET}$. From the two plots, we can observe that programming to a higher $R_L$ requires less time and energy than a lower $R_L$. Also, a larger $R_L$ can suppress the leakage current flowing through unselected cells. However, to maintain sufficient read margin, a smaller $R_L$ is preferred, i.e., larger $R_H/R_L$. Although a cell can be written by low voltage, the long switching time makes write operation less energy-efficient. These data provide us some clues to investigate the optimal block size and operation voltage during write and read operation.

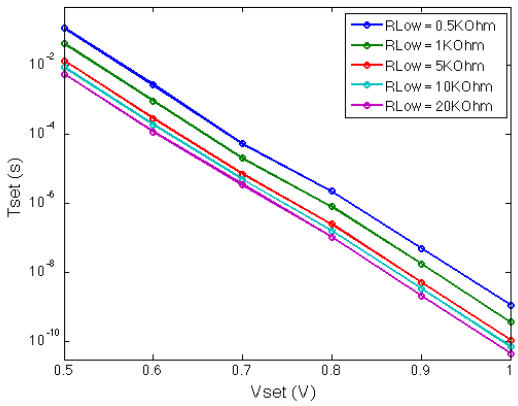



Fig. 6. Write time of one cell under different VSET and RL.

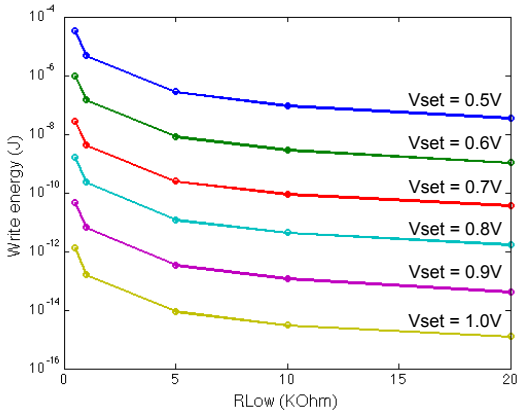



Fig. 7. Write energy of one cell under different VSET and RL.

### A. Write operation

For crosspoint structure, we analyzed four possible schemes for write operation [9]: *HWHB* activates the selected wordline (WL) and bitline (BL), and half biases unselected WLs and BLs; *FWFB* activates the selected WL and BL, and leaves unselected WLs and BLs floating; *HWFB* (*FWHB*) activates the selected WL and BL, half biases unselected WLs (BLs), and leaves unselected BLs (WLs) floating.

To minimize energy consumption during write operation, three schemes (take *HWFB* same as *FWHB*) are compared in terms of energy efficiency. Fig. 8 illustrates the leakage path in three schemes, by which we can estimate the cell current of the selected cell and leakage current of all unselected cells. The calculated leakage currents under the worst case (all unselected cells are $R_L$) are shown in (1) – (3).

$$HWHB: \quad I_{leakage} = \frac{V_{SET}}{2} \times \left(\frac{n-1}{R_L} + \frac{m-1}{R_L}\right) \tag{1}$$

$$FWHB: \quad I_{leakage} = \frac{V_{SET}}{2} \cdot \frac{n-1}{R_L} \tag{2}$$

$$FWFB: \quad I_{leakage} = V_{SET} \cdot \frac{(n-1)(m-1)}{(n+m-1) \cdot R_L} \tag{3}$$

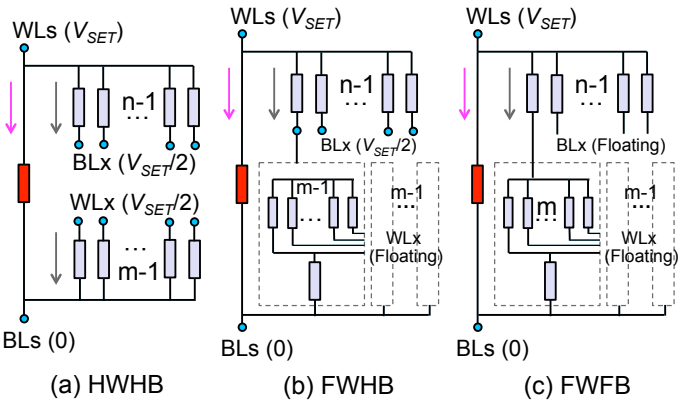
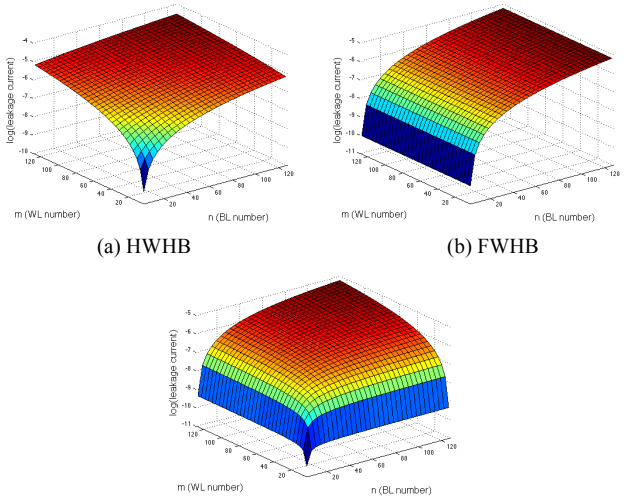



Fig. 8. Current path through selected and unselected cells in (a) *HWHB*, (b) *FWHB*, and (c) *FWFB*.

Fig. 9 shows the total leakage current in three different schemes. It is obvious that *HWHB* consumes the most leakage

current, while *FWHB* and *FWFB* flow similar amount of leakage current when BL number equals WL number. However, *FWFB* has an inherent problem that may result in write disturb. Floating both unselected BL and unselected WL may lead to more than $V_{SET}/2$ drop on an unselected $R_H$ and switch it to $R_L$. Therefore, we will use *FWHB* scheme in our designs.




(a) HWHB (b) FWHB

(c) FWFB

Fig. 9. Total leakage currents in three schemes. ($V_{SET}$ = 0.9V, $R_L$=10KΩ)

### B. Read operation

A simple and instinctive way to read out the cell state is given a fixed voltage and measuring the current. In the parallel read scheme, BL voltage variation due to data pattern results in leakage path, which degrades the read margin. Therefore, while designing current-mode sensing amplifier, one key point is to minimize the undesired leakage current, which is a tradeoff between transistor size and BL variation in diode-connected current sensing scheme. Also, voltage/process variation and wide cell distribution make the sensing amplifier design more challenging.

Since it is still difficult to thoroughly eliminate the leakage path due to small variations, block size is limited for the worst-case scenario in read operation. The worst case happens when sensing a high resistance as all the other cells are in low resistance state. In this case, the neighbor BL would have slightly higher voltage than the BL with $R_H$ cell, as shown in Fig. 10. To calculate the maximum unit block size, we should make sure the leakage current is smaller than the cell current.

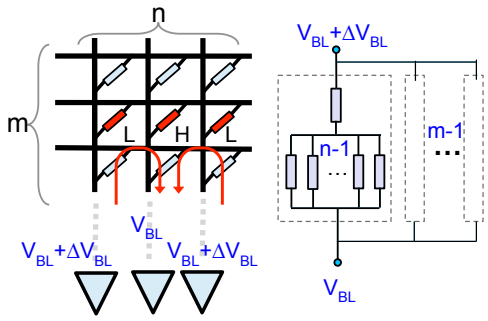



Fig. 10. Worst-case scenario of parallel read and the equivalent leakage path.

## IV. Peripheral Circuit

Peripheral circuit design includes WL/BL drivers, sensing amplifiers, control circuits, write voltage generator and power management to every unit block. We are not discussing how to generate $V_{SET}$ and $V_{RESET}$ at this point, since we would like the programming voltages be flexible. WL/BL drivers take the decoded address signals and data-in value to provide correspondent voltage and sufficient current for selected and unselected WLs/BLs. Sensing amplifiers take the difference of the selected cell and the reference cell in terms of voltage or current, amplify the small analog difference and output a digital code. It highly depends on the cell distribution to successfully read out the right data. Both WL/BL drivers and sensing amplifiers should be designed as small as possible to fit in the narrow cell pitch. Control circuit computes all the input signals to determine the current operation state and output internal control signals to other circuits. Power distributed to each unit block should be carefully handled while enabling the accessed unit block and disabling the other idle ones to minimize leakage current.

## V. Future Work

In the future weeks, we will mainly put our effort on building up a whole RRAM circuit, including crosspoint array and peripheral circuit. In the meantime, some circuit techniques will be explored, analyzed and compared. Cell distribution is also an important factor that affect array yield. How to suppress the effect of wide cell distribution is a key in peripheral circuit design. Finally, we would like to investigate the possibility of RRAM as a cache. In this case, typical SRAM features need to be characterized and compared with RRAM, especially the energy consumption during standby mode. From this analysis we can determine what application is the best for RRAM to replace SRAM as a cache.

## References

[1] Thatcher Jonathan, et al., "NAND Flash Solid State Storage for the Enterprise, An in-depth Look at Reliability," *Solid State Storage Initiative (SSSI) of the Storage Network Industry Association (SNIA)*, April 2009.

[2] Elaine Ou and S. Simon 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.

[3] ITRS Roadmap (http://www.itri.net)

[4] 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.

[5] R. E. Simpson, et al., "Toward the Ultimate Limit of Phase Change in $Ge_2Sb_2Te_5$," *Nano Letter*, pp. 414-419, 2010.

[6] 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.

[7] Yi-Chou Chen, et al., "An Access-Transistor-Free (0T/1R) Non-Volatile Resistance Random Access Memory (RRAM) Using a Novel Threshold Switching, Self-Rectifying Chalcogenide Device", *IEEE IEDM*, pp. 37.4.1-37.4.4, 2003.

[8] Cong Xu, et al., 'Design Implications of Memristor-Based RRAM Cross-Point Structures', DATA, pp. 1-6, 2011.

[9] D. Niu, C. Xu, N. Muralimanohar, N. P. Jouppi, Y. Xie, "Design Trade-Offs for High Density Cross-Point Resistive Memory," *ISLPED*, 2012, pp. 209-214.