A Variation-Tolerant Multi-Level Memory Architecture Encoded in Two-state Memristors

Bin Wu and Matthew R. Guthaus
Department of CE, University of California Santa Cruz
Santa Cruz, CA 95064
{wubin6666,mrg}@soe.ucsc.edu

Abstract

Memristors are becoming a promising non-CMOS high-density memory solution as CMOS technology approaches atomic limits. However, high electrical variability of both memristors and the analog reading circuitries cause significant error rates and the use of transistors limits the density of memristor/transistor hybrid architectures. This work presents a multi-memristor cell design that is robust while retaining the simplicity of non-feedback memristor programming. The proposed architecture offers a high bit density compared to other memristor/transistor hybrid architectures by introducing multilevel outputs to store multiple bits per cell, and has competitive power and read speed to existing architectures.

I. Introduction

Memristors, theorized many years ago [1] and recently discovered as passive devices [2], are a promising non-CMOS memory technology with the potential for very high density, low stand-by and refresh power, non-volatility, and a long lifetime [2]. The physical structure of memristors is a metal-insulator-metal (MIM) structure, which consists of three layers: two metal layers as interfaces with metal wires, and one oxide material layer in between as an insulator.

One advantage of memristor memory is the potential for high density. Single memristors can be 10nm or less in horizontal dimensions [3]. The two states of the memristor, high resistance state (HRS) and low resistance state (LRS), can be used to store binary information. Memristors can also be built "above" a transistor substrate [4] in the interconnect stack to work with existing CMOS technology.

Resistance variation is the main challenge designing with memristors. Such variation is caused by numerous factors such as the random nature of the conductive filament forming process, device area, programming current amplitude and programming duration. Programmed LRS/HRS resistance varies not only from device to device, but also from cycle to cycle [5], [6]. In addition, process variation causes mismatch in analog circuitries that read the memristor current. All of these factors directly influence the output value and the variation tolerance of different architectures is inconsistent.

The first design decision is whether to use transistors in the memory cell or only memristors. The well-known crossbar structure doesn't use any transistors by combining horizontal and vertical select lines to isolate a single memory cell [2]. On the other hand, a memristor/transistor hybrid design similar to a DRAM can use an access transistor for each memristive device (1T1M) to isolate a memory cell.

Sneak current is the main disadvantage of memristor memories without transistors. In the example of the crossbar, sneak currents flow through unselected memristors in LRS and cause potential errors. Operational amplifiers (op-amps) and nonlinear memristors are the two possible solutions. Op-amps can hold the output node at a fixed reference voltage to eliminate voltage difference between the terminals of unselected memristors, but costs additional power and area [7]. Nonlinear I-V behavior of memristors within a single state, especially LRS, can also alleviate this issue, but this requires mature device engineering which is still evolving [8].

This work focuses on using hybrid memristor/transistor designs, which eliminate sneak current by using the transistors to isolate unselected cells. However, transistors, which are bigger than memristors, can limit the cell density.

The second design decision is the number of memristors used per cell. The crossbar structure and 1T1M are single memristor (SM) cells. Multiple memristors (MM) can be used to build a cell that has high device variation tolerance over a SM cell, and still do not require extra analog reading circuits, avoiding another potential variation source.

The third design decision is the number of output levels. Most common memory cells contain one bit and generate a two-level output, but increasing the number of output levels can increase the bit density. Multi-level outputs do not necessarily require multi-memristor cells as a single memristor can have multiple resistive states besides just LRS and HRS.

Combining the above choices, there are four possible memristor memory cell styles: Single-Memristor 2-Level (1M-2L), Multi-Memristor 2-Level (MM-2L), Single-Memristor Multi-Level (1M-ML), and Multi-Memristor Multi-Level (MM-ML). This work proposes the first 2-memristor 3-level (2M-3L) cell design that is a compromise between architecture complexity, bit density, and variation tolerance. In particular, this work contributes:

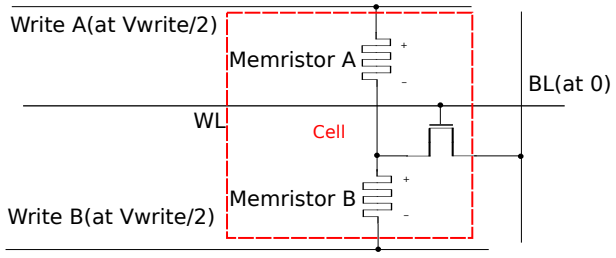
- The first variation-tolerant 2M-3L cell memory architecture.
- New variability and density studies of previous 2M-2L cell memory architectures.

The rest of the paper is organized as follows: Section III introduces proposed 2M-3L cell. Section IV analyzes performance and Section VI present our results. Section VII concludes the paper.

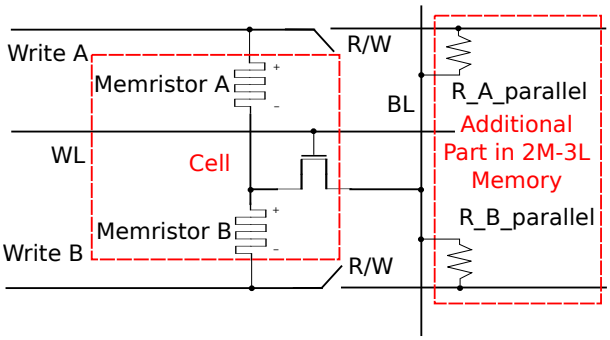
II. Previous Work

The design of a 2M-2L cell is similar to a complementary CMOS inverter except with a pull-up and pull-down memristor in place of transistors as shown in Figure 1(a) [9] . The output is accessible through an NMOS access transistor connected to the internal node and a shared bit-line, BL.

During a write operation, the positive (negative) voltage difference between $Write_A$ and BL writes 0 (1) to memristor A, while a similar operation with $Write_B$ applies to memristor B. During a read operation, the read voltage is applied between $Write_A$ and $Write_B$, and the resistance ratio of the two devices determines the voltage output on BL.



(a) One variant of multi-memristor cells uses a pull-up and pull-down memristor much like a CMOS inverter [9].



(b) By adding two resistive loads per bit line, 2M-3L cell can generate one extra logic level (HRS-HRS) compared to 2M-2L cell.

Fig. 1. Multiple-memristor cells (MM-C) use two (or more) devices to improve noise margins along with an access transistor.

III. 2M-3L cell Design

The design of a 2M-2L cell is similar to a complementary CMOS inverter except with a pull-up and pull-down memristor in place of transistors [9]. The output is accessible through an NMOS access transistor connected to the internal node and a shared bit-line, BL. Typically, only two complementary states of the two devices are used: LRS-HRS and HRS-LRS.

However, a complementary memristor cell has four possible binary states of the two devices which enables two bits to be stored in each memory cell. Using all four states requires a cell modification to block the LRS-LRS short-circuit path and requires higher voltage resolution during reading. Hence, we propose a 2M-3L cell, instead of a 4-level cell to avoid this.

The architecture of a 2M-3L cell is similar to a 2M-2L cell, but it use two resistors per bit line in parallel to generate the additional HRS-HRS output as illustrated in Figure 2. These Parallel Resistors (PR) have a negligible effect on the density as they are shared per bit-line. The PR resistance is smaller than the HRS resistance to suppress the HRS variation effect on the output, but it is much larger than LRS resistance so that LRS memristors can still short circuit a PR when in parallel, and to avoid short circuit from supply to ground.

During a read operation, the combination of memristor A in LRS and memristor B in HRS is defined as the LRS-HRS state. The LRS-HRS state has a low resistance between $Write_A$ and the output node, and thus can produce a high logic output due to the high PR/LRS ratio. Similarly, an HRS-LRS state will produce a low logic output. An HRS-HRS state, however, will

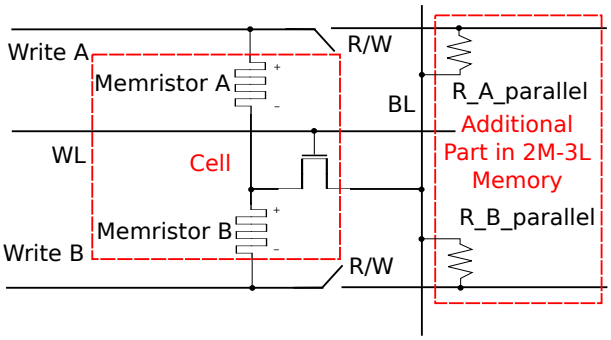


Fig. 2. Multiple-memristor cells use two (or more) memristors with an access transistor while our proposed scheme uses Parallel Resistors (PR) shared among an entire bit-line to read multiple output levels.

generate a middle logic value. Since PR is smaller than HRS and is in parallel with both HRS devices, HRS variation has very little affect on the output voltage.

The write operation uses the bit lines and both write lines. The voltage difference on the two write lines can write the whole row of cells to an HRS-HRS value. If another value needs to be programmed, the two write lines must stay at the same voltage and the bit line will be higher or lower than that voltage. The write operation needs more than one stage if LRS resistance is not much larger than the transistor ON resistance. The programmed LRS device will significantly reduce the voltage applied to the two memristors, and sabotage the voltage required to further increase the resistance of the other memristor, which needs to become HRS. The first write step is to initialize all the cells to be written to HRS-HRS with following programming voltage: $Write_B$ at $2\times V_{write}$, BL at V_{write} and $Write_A$ at 0. If HRS-HRS is the desired state, the writing is done for the cell. Otherwise, the second stage is to program one HRS device to LRS. Both $Write_A$ and $Write_B$ are at V_{write}, the BL will be $2\times V_{write}$ if memristor A is to be made LRS or 0 if memristor B is to be made LRS.

Ternary encoding and decoding cost add little extra power and area, since the costs can be amortized over an entire memory array. The truth table of the encoder and decoder circuits are shown in Table I. Implemented with digital, dynamic logic, these digital auxiliary circuitries are fairly robust to variation compared to analog circuits. Each encoder/decoder circuit handles two 2M-3L cells which represent 3 bits of information. The encoder circuit takes a 3-bit input and generates a 4-bit output to program two 2M-3L cells, while the decoder circuit takes a 4-bit input to generate a 3-bit output. Define the encoder output bits as $M_3M_2M_1M_0$ and the decoder output bits as $B_2B_1B_0$, then the the output expressions are: $M_3 = 101 + 0xx$, $M_2 = 1xx + 01x$, $M_1 = 0xx + 1x0$, $M_0 = 11x + 1x1 + 0x1$, $B_2 = 11xx + xx11$, $B_1 = 111x + 1x10$, and $B_0 = x01x + xx11$. Inverters with different threshold voltages are used to distinguish a high, middle or low output. One approach of changing the inverter threshold voltage is to set different bias voltages. Threshold voltage variation can be suppressed by using large transistors since they are amortized.

TABLE I

TRUTH TABLE OF ENCODER AND DECODER CIRCUITRIES, THEY CAN BE IMPLEMENTED BY USING DYNAMIC LOGIC AND MULTI-THRESHOLD INVERTERS.

Device State	Ternary Bits	Input Bits	Encoder Output	Inverters Outputs	Decoder Outputs
L-H&L-H	HH	111	0101	1111	111
L-H&H-H	HM	110	0111	1110	110
H-H&L-H	MH	101	1101	1101	101
L-H&H-L	HL	100	0110	1100	100
H-L&L-H	LH	unused	-	0011	unused
H-H&H-H	MM	011	1111	1010	011
H-H&H-L	ML	010	0111	1000	010
L-H&H-H	LM	001	0111	0010	001
H-L&H-L	LL	000	1010	0000	000

IV. PERFORMANCE COMPARISON

The density advantage of 2M-3L over 2M-2L depends on the technology size ratio of a transistor to the memristor. If the ratio is over 2, the 2M-3L cell bit density is 50% higher compared to the 2M-2L cell bit density, as two 2-level cells can replace three 2-level cells. This is the most likely situation as the memristor structure is similar to an interconnect via. The density improvement will be less 50% if the ratio is smaller 2 and can be negative when the ratio is less than 1. Conversion circuits also require some area overhead.

The main advantage of MM cells is the high variation tolerance compared to SM cells. If we define x as the relative deviation of a programmed memristor value to the nominal one, the effect of variation can be formalized as shown in Table II. In 1M-2L memories, a reference resistance should be defined to distinguish HRS and LRS, and is written as R_{ref} in Equation (1). It should be the equivalent resistance to generate the threshold current between LRS and HRS current in the reading circuitries. R_{mem} is the resistance of the memristor, and can be at HRS or LRS. In 2M-2L memories, such reference of a memristor is the other memristor in the cell. R_A and R_B in Equation (3) are the two memristor resistance values. One of these two is at HRS while the other one is at LRS. In 2M-3L memories, the two parallel resistors, written as R_p, are a reference. In Equations (3) and (5), R_t is the transistor drain to source resistance, R_A and R_B are the two memristor resistance values, and R_p is the resistance of the parallel resistors. An x of 0.8 can program the value to be $0.2\times$ or $1.8\times$ of the expected value. The reported memristor variation is around 10 times [9], hence, the upper limit of $|x|$ should be around 0.8 and is much smaller than C_2. The worst case for all memories is when the coefficient C_1 in denominator is less than one and the constant C_2 dominates the values of the derivative. The upper limits of Equations (4) and (6) are smaller than 0.25, while the upper limit of Equation (2)

TABLE II

OUTPUT AND VARIATION DERIVATIVE FUNCTIONS SHOW THAT 2M CELLS ARE MORE ROBUST TO DEVICE VARIATION THAN 1M CELL

1M-2L cell	
V_{out}	$\frac{V_{read}}{R_{mem}(1+x)}(\frac{V_{read}}{R_{ref}})^{-1} = \frac{1}{\frac{R_{mem}}{R_{ref}}(x+1)}$ (1)
$\frac{dV_{out}}{dx}$	$\frac{1}{(C_1x+C_2)^2}, C_1 = \sqrt{\frac{R_{mem}}{R_{ref}}}, C_2 = 1$ (2)
2M-2L cell	
V_{out}	$\frac{V_{read}R_B}{R_A(1+x)+R_B}{V_{read}}^{-1} = \frac{1}{\frac{R_A}{R_B}(1+x)+1}$ (3)
$\frac{dV_{out}}{dx}$ of Memristor A	$\frac{1}{(C_1x+C_2)^2}, C_1 = \sqrt{\frac{R_A}{R_B}}, C_2 = \frac{R_A+R_B}{\sqrt{R_AR_B}}$ (4)
2M-3L cell	
V_{out}	$\frac{\frac{1}{R_A(1+x)}(R_t+R_p)+\frac{R_t}{R_B}+1}{(2R_t+R_p)(\frac{1}{R_A(1+x)}+\frac{1}{R_B})+2}$ (5)
$\frac{dV_{out}}{dx}$ of Memristor A	$\frac{1}{(C_1x+C_2)^2}$ (6) $C_1 = \sqrt{\frac{R_A}{R_BR_p}}\frac{2R_B+R_p+2R_t}{\sqrt{R_B+R_p+2R_t}} < \sqrt{\frac{R_A}{R_B}}$ $C_2 = (\frac{\sqrt{R_A}}{\sqrt{R_B}}+\frac{\sqrt{R_B}}{\sqrt{R_A}})\sqrt{\frac{R_A+R_p+2R_t}{R_p}} > 2$

is 1. The difference between Equations (3) and (5) varies, but the noise margin for a 2M-3L cell is much smaller than a 2M-2L cell due to the extra output level.

The read speed of a 2M-3L cell depends on the status of the cell. An HRS-HRS cell delivers a very small current, while HRS-LRS and LRS-HRS cells deliver enough current to achieve fast read speeds. The slow HRS-HRS read speed can be eliminated by pre-charging the BL to V_{read}/2. The cells do not to change the bit line when the cell status is HRS-HRS. The write speed of 2M-3L cells can be longer due to more programming stages compared to single device programming.

Read power per bit is improved compared to that of 2M-2L cells. Reading a 2M-3L cell requires the same amount of energy to read a 2M-2L cell, but provides more information.

Theoretical write power performance can be calculated assuming all designs possess the same device write power. The best 2M-2L cell write power is $2\times$ the 1M-2L cell write power. The 2M-3L cell write power depends on the write pattern. Changing the cell between HRS-LRS and LRS-HRS requires two device writes while the rest request one. The expect 2M-3L cell write power is $\frac{4}{3}\times$ a 1M-2L cell assuming a random input data pattern, and the write power per bit is $\frac{4}{3}\times\frac{2}{3} = 0.89\times$ the 1M-2L cell write power, as every two 2M-3L cells store three bits. There are several factors that can increase the 2M-3L write power above the theoretical minimum, however. The extra reset cycle increases the write power of 2M-3L cell slightly as the HRS devices only conduct off current. The encoder/decoder power consumption is not considered, but its influence can be negligible since the device programming contributes most of the writing power. The simulation results verify this.

V. SIMULATION SETUP

All the architectures are assumed to be fabricated with a 45nm technology memory compiler [10]. Due to the absence of a physical memristor model, a metal 1 to metal 2 via with a side length of 70nm, is used to represent a memristor because of the similar physical structures. The memristor spice model is based on an existing work [11] but with several modifications. HRS/LRS is set at $2M\Omega$/$2K\Omega$, which is within reported memristor resistance value range [12], to ensure the read current per bit is at the mA level to avoid chip power budget problems. Ideal switches are added in the memristor model so that the

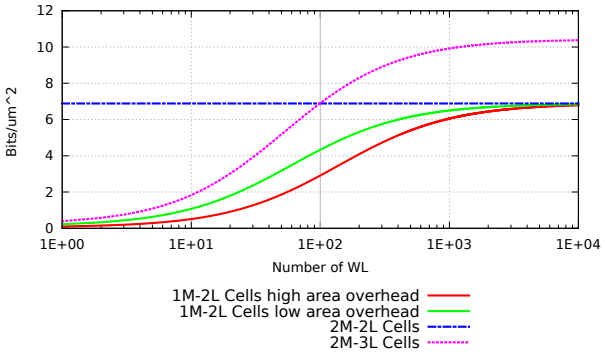


Fig. 3. 2M-3L memories density is the highest after word line size of 100.

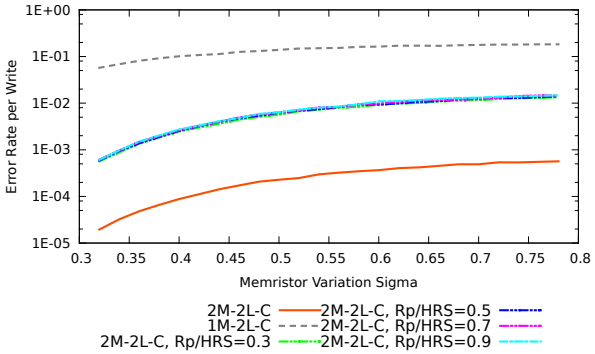


Fig. 4. 2M-3L cells sacrifice some variation tolerance compared to 2M-2L cells but is much more robust than 1M-2L cells.

devices exhibit saturated charge characteristics, which ensures the time to fully program a device is not related to previous programming time and will remain constant for one type of memristor.

A column of cells that share the same bit line are used to simulate the read operation and a row of cells that share the same word line are used in the write simulation. Unused memristors are replaced by resistors to eliminate the impact of write operation quality. Monte Carlo simulations are used to analyze the noise margins using the analytical formulations in Equations (1), (3) and (5). This is because the error rates per write are quite low and several millions sample cases are needed to ensure accuracy. Each error rate data is calculated when 2000 error cases are collected to ensure the result is repeatable. Memristor variation is simulated by adding Gaussian variation to the memristor resistance.

VI. Results

Fig 3 shows that 2M-3L memory density is highest for larger arrays. 1M-2L memories has similar initial area overhead but lower bit density per cell. 2M-2L memory density is a constant. 1M-2L memories can have reading circuit areas per bit line that range from $8.64um^2$ to $20um^2$ [13], [14]. Decoder and encoder circuits area per bit line are $12um^2$ in 2M-3L memories.

Compared to 2M-2L memories, MM-3L memories sacrifice some device variation tolerance for improvement in density, but still show better tolerance than 1M-2L memories. Smaller parallel resistors (R_P) show limited improvement in control over device variation. A design with $\frac{R_P}{HRS}$ ratio of 0.99 only shows a $2\times$ error rate of a design with that ratio at 0.01.

Simulated power and speed results are shown in Table III. The cost of variation tolerance in 2M-2L memories is a disadvantage in read speed, write power and write speed. By using memristors more efficiently, the 2M-3L cell shows improvement over 2M-2L cell except write delay.

Compared to 2M-2L memories, 2M-3L memories increase density by 50%, reduce read power by 40% per bit and and write power by 60% with some initial cost. Only a small sacrifice is made in variation tolerance (1% error rate over a 1M-2L memory). Read speed remains the same per word line. The worst writing delay can be $2\times$ of 2M-2L memory write delay depending on the write pattern.

VII. Conclusion

The proposed 2M-3L memory shows good variation tolerance compared to 1M-2L memories and higher density over the other both single-level designs. The costs are longer write delay and potentially higher read delay with large array size. Write delay can be mitigated by using a cache to buffer the writes since these are often not performance critical.

TABLE III

2M-3L MEMORIES SHOWS BETTER PERFORMANCE OVER 2M-2L MEMORIES BESIDES WRITING DELAY. RESULTS BETWEEN 2M-3L MEMORIES AND 1M-2L MEMORIES VARY ON MEMORY CONFIGURATIONS.

Memory	1M-2L	2M-2L	2M-3L
Simulated Write Speed (ns)	10	13	26
Simulated Read Speed (ps)	SA Speed	20/WL	20/WL
Theoretical Write Power per bit	1	2	0.89
Simulated Write Power per bit (pJ)	0.07	0.27	0.12+encoder(.002)
Theoretical Read Power per bit	P_{SA}	1	0.66
Simulated Read Power per bit (aJ)	43.8/WL +P_{SA}(10e5 level)	43.8/WL	31.1/WL+ decoder(22k)

REFERENCES

[1] L. O. Chua, "Memristor-the missing circuit element," *Circuit Theory, IEEE Transactions on*, vol. 18, no. 5, pp. 507–519, Sep 1971.

[2] R. S. Williams, "How we found the missing memristor," *Applied Physics Letters*, vol. 45, 2008.

[3] B.Govoreanu, G.S.Kar, Y-Y.Chen, *et al.*, "10x10nm2 Hf/HfO_x crossbar resistive RAM with excellent performance, reliability and low-energy operation," *IEDM*, 2011.

[4] I. G. Baek, M. S. Lee, S. Seo, *et al.*, "Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses," *IEDM Tech. Dig.*, vol. 13, pp. 587–590, 2004.

[5] B. J. Choi, A. C. Torrezan, K. J. Norris, *et al.*, "Electrical performance and scalability of pt dispersed SiO_2 nanometallic resistance switch," *Nano Lett.*, 2013.

[6] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-based resistive switching memories nanoionic mechanisms, prospects, and challenges," *Advanced Materials*, 2009.

[7] M. Qureshi, M. Pickett, F. M. F., *et al.*, "Cmos interface circuits for reading and writing memristor crossbar array," *ISCAS*, 2011.

[8] J. J. Yang, M.-X. Zhang, M. D. Pickett, *et al.*, "Engineering nonlinearity into memristors for passive crossbar applications," *Applied Physics Letters*, vol. 100, 2012.

[9] D. Sacchetto, P.-E. Gaillardon, M. Zervas, *et al.*, "Applications of multi-terminal memristive devices: A review," *IEEE Circuits and Syst. Mag.*, vol. 13, 2013.

[10] M. Guthaus, "Openram: An open-source memory compiler," *ICCAD*, 2016.

[11] D. Biolek, M. D. Ventra, and Y. V. Pershin, "Reliable SPICE simulations of memristors, memcapacitors and meminductors," *Radioengineering*, vol. 22, no. 4, pp. 945–968, 2013.

[12] M.-S. Wong, H. Lee, S. Yu, *et al.*, "Metal-Oxide RRAM," *Proc. IEEE*, 2012.

[13] A.-T. Do, Z.-H. Kong, K.-S. Yeo, *et al.*, "Design and sensitivity analysis of a new current-mode sense amplifier for low-power SRAM," *TVLSI*, 2011.

[14] S.-S. Sheu, M.-F. Chang, K.-F. Lin, *et al.*, "A 4Mb embedded SLC resistive-RAM macro with 7.2ns read-write random-access time and 160ns MLC-access capability," *TVLSI*, 2011.