Extended Abstracts of the 2019 International Conference on Solid State Devices and Materials, Nagoya, 2019, pp693-694

Novel Dot-Product Engine for Low-Voltage Analog Oxide ReRAM Array

Hasita Veluri, Yida Li, Maheswari Sivan, Jessie Xuhua Niu, Umesh Chand, Jin Feng Leong, Evgeny Zamburg, Subhranu Samanta, Xiao Gong, Kelvin Xuanyao Fong, and Aaron Voon-Yew Thean
National University of Singapore, Dept. of Electrical & Computer Engineering, 4 Engineering Drive 3, Singapore 117583
Email contact: li.yida@nus.edu.sg, Aaron.Thean@nus.edu.sg

Abstract: **In this paper, we demonstrate a novel dot product engine for in-memory matrix-matrix (M2M) multiplication using SiO_2 based resistive random access memory (ReRAM). The CMOS compatible ReRAM exhibits switching at sub 1.2V, and is fabricated using a low-temperature process, thus making it suitable for multilayer 3D sequential memory integration in the BEOL of CMOS technologies. Our proposed M2M method uses non-linear ReRAMs and requires only a 4-bit digital-to-analog converter (DAC). In addition, we implemented operations for positive and negative numbers in the same array. As a result, we achieve 33% power reduction and 40% area reduction as compared to conventional M2M.**

1. Introduction

Despite sneak current issues, 1-R cross-point ReRAM array architectures have shown to be viable for in-memory computing such as full adder and matrix multiplications [1-2]. However, there are issues in existing implementations that need to be addressed. Firstly, the conventional method of mapping elements of a matrix to different read voltage levels cannot be easily applied for non-linear ReRAMs. Secondly, M2M using binarized inputs, suitable for neural networks, results in loss of accuracy for dot product operations for accurate floating point number computations.

To accommodate non-linear ReRAMs, we developed a M2M algorithm that maps matrix elements to pulse-number based operations instead of analog voltages levels. The algorithm performs operations on positive and negative numbers using the same array and utilizes a 4-bit DAC. Our implementation reduces power consumption and area by 33% and 40% respectively, as compared to conventional M2M [3]. We demonstrate our approach for a low-voltage SiO_2-based ReRAM array (24x24) that exhibits sub-1.2V switching voltages. We expect the combination of ReRAM technology and system algorithm innovations to dramatically lower power consumption of in-memory computing.

2. Device Fabrication and Characterization

The CMOS-compatible W-SiO_2-Ti ReRAM process flow is detailed in **Fig. 1**. The DC characteristics are shown in **Fig. 2**. The first positive sweep forms the device at ~1.7 V, and subsequent set and reset voltages (V_{set} and V_{reset}) fall in the sub-1.2 V region. The ReRAM exhibits bipolar resistive switching behavior due to the active Ti top electrode (TE) metal that modifies the distribution of the oxygen vacancies in SiO_2. The formation and rupture of conducting filament likely happens at the Ti-SiO_2 interface, leading to the observed analog behavior (gradual set/reset) [4].

Figs. 3 and 4 show LRS/HRS up to 10^4 Read/Write cycles (set/reset pulse of 0.5/-1V), and the retention time respectively. Reasonable stability of the switching characteristics can be seen, with wider variations in the HRS state. **Fig. 5** shows the spread of LRS/HRS of 15 devices, showing larger variation at the HRS. **Fig.6** shows the conductance of the ReRAM at a read voltage of 0.6 V over 75 reset pulses applied. Device is set to LRS at 1.6 V, followed by reset voltage pulses of -1 V and 300 ns width. The trend of reducing conductance as a function of pulse number is clearly seen, demonstrating the potential for analog programming. **Fig. 7** shows the benchmark table comparing the switching characteristics of the SiO_2 ReRAM in this work with other reported SiO_2-based ReRAM cells. The device reported in this work exhibits the lowest V_{set}/V_{reset} and process thermal budget.

3. Matrix-Matrix Multiplication Algorithm

Fig. 8 illustrates the M2M methodology in this work. Values of matrix A are first mapped onto the ReRAM conductance states through a state transformation as illustrated in **Fig.9**. Since ReRAM conductance is not linear with applied voltage **(Fig 2)**, Matrix B elements are mapped to read pulse numbers. Bit line (column) currents are integrated over the read time and converted to V_{out}. This is inverse transformed to return the floating-point result. In the state transformation, each matrix is expressed as the sum of a unit matrix multiplied by its minimum value and a residual matrix which normalizes all the values in the residual matrix to be ≥ 0. Thus all-positive operations can be performed with the array for both positive and negative numbers while avoiding the use of additional arrays, control circuitry as reported [3]. **Fig. 9** shows a detailed flow chart of the developed M2M algorithm.

In our proposed method, to minimize error due to the intrinsic ReRAM variability, we chose to quantize our states close to the LRS (large variability for states close to HRS) **(Fig. 6).** A compact ReRAM model reported by [5] has been used to fit the device electrical characteristics, allowing us to evaluate the effects of increasing matrix dimensions and resolution on output error. **Fig. 10** shows that average output error for lower dimension matrices is less than 20% and as matrix dimension increases the error drops to less than 3%. In addition, with greater ReRAM states, the average error can be reduced to less than 10% even for lower dimension matrices. **Fig. 11** shows an illustration of the control system built to implement the algorithm for matrices with dimensions $\leq$ (6x6), on the ReRAM array. A good agreement between simulated and experimental data was obtained. **Fig. 12** compares the performance parameters – energy, area and computation time with conventional M2M. Our proposed algorithm reduces total power consumed by 33% and area by 40%, while resulting in no increase in computation time.

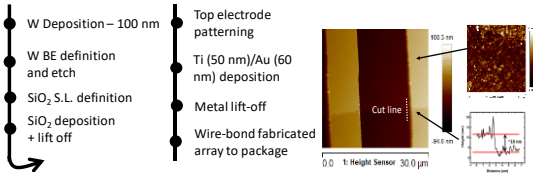
4. Conclusion

We demonstrated a dot product engine for floating point number matrix multiplication using non-linear sub-1.2V-based analog oxide ReRAM array. Our approach minimizes quantization error arising from mapping of positive and negative numbers to the discrete conductance states of the non-linear ReRAM. The paired technology-system approach achieves 33% power and 40% area reduction, as compared to conventional M2M.

Acknowledgements
The work is supported in part by Singapore's National Research Foundation (NRF-RSS2015-003), A*Star AME Grant A1892b0026, Hybrid Integrated Flexible Electronic Systems (HiFES) Program (hifes.nus.edu.sg) and E6Nanofab at the National University of Singapore (NUS).

References

[1] B. Chen et al, 2015 IEDM Conference ; [2] Y. Liao et al, 2018 VLSI Symposium; [3] L. Xia. et al., 2016 J.Comput.Sci. Technol.; [4] B.Gao et al, 2009 IEEE Elec. Dev. Lett.; [5] E.Lehtonen et al, 2010 CNNA



(a) (b)

Fig. 1 (a) Process flow showing the fabrication of W-SiO_2-Ti ReRAM. (b) AFM characterization of the SiO_2 switching layer. The SiO_2 thickness and roughness above the bottom electrode is ~12 nm and ~10 Å respectively. (W) as the bottom electrode (BE) and titanium (Ti) as the top electrode (TE).

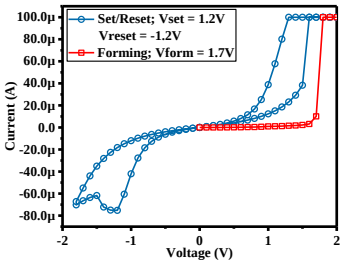


Fig. 2 DC curve of the SiO_2 ReRAM. The 1st cycle forms device at ~1.7V. V_{set}=1.2V and V_{reset}=-1.2V

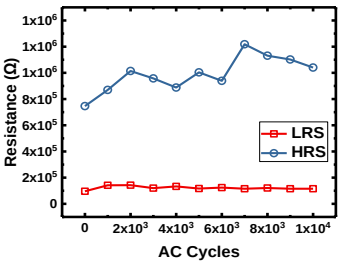


Fig. 3 Endurance characteristics at a read voltage of 0.1V. Wider variations in HRS resistance.

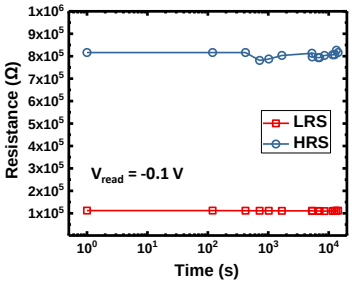


Fig. 4 Retention behavior of both resistance states measured for W-SiO_2-Ti ReRAM. V_{read} used is -0.1 V.

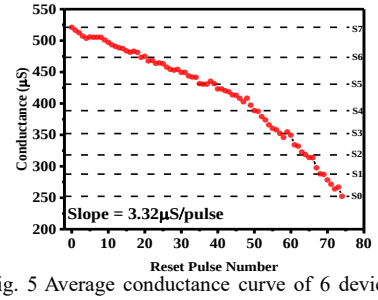


Fig. 5 Average conductance curve of 6 devices VS reset pulses. V_{reset} = -1V, V_{read} = 0.6V, T_{reset} =300ns, T_{read} =30ns. V_{set} =1.4V, CC=1mA. Conductance decreases by 2x over 75 reset pulses and is divided into 8 states.

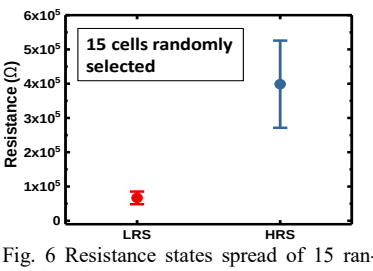


Fig. 6 Resistance states spread of 15 randomly selected devices. V_{read} = -0.1V, resistance extracted from DC curves. Spread of HRS is much higher than LRS spread.

Device Stack	Thickness (nm)/ Temperature (°C)	V_{set} (V)	V_{reset} (V)	Retention (s)	AC Endurance cycles (N)
W/SiO_2/Ti (This work)	12/Room temp	< 1.2	< 1.2	$>10^4$	$>10^4$
TiN/SiO_2/p-Si	4/600	1.1-1.4	-1.4-1.6	$>10^4$	$>10^4$
Au/Zr/SiO_x/n+Si	40/400	3.5	3	$>10^4$	$>10^4$
p+Si/SiO_2/n+Si	5/Room temp	7	5	>2 x 10^5	N/A
Au/Cr/SiO_2:W/Si	25/1000	1.5-2.5	-0.5-1.5	$>10^4$	400
Nanosphere/SiO_2/Si	40/Room temp	3-3.4	5-8	10^4	N/A

Fig. 7 Benchmark table comparing the resistive switching characteristics of SiO_2-based ReRAM cells. The W-SiO_2-Ti ReRAM reported in this work exhibits the lowest V_{set}/V_{reset} as well as the processing temperature.

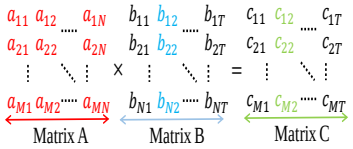


Fig. 8 In-Memory Matrix-Matrix Multiplication method. Matrix A mapped to ReRAM conductance, Matrix B mapped to number of read pulses. $C = A \times B$

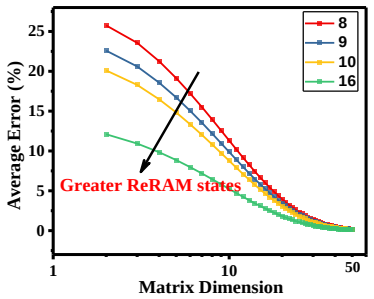


Fig 10. Maximum average error VS ReRAM states and matrix dimension. Maximum error decreases to <3% with increase in dimension and to <10% for higher ReRAM states at lower dimensions.

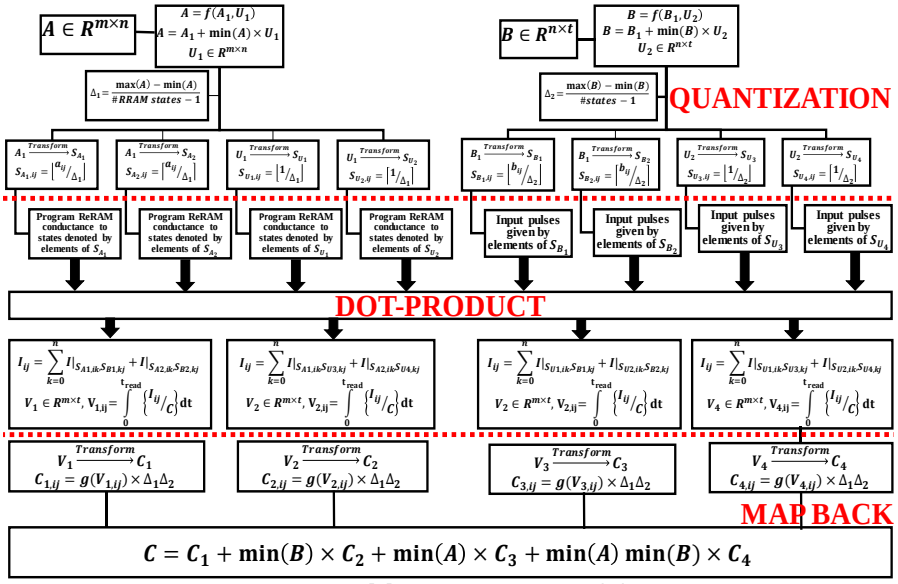


Fig. 9 Flowchart for developed algorithm.$\lceil x \rceil$: ceiling function of x; $\lfloor x \rfloor$: floor function of x; $U_{1,2}$: unit matrices; A,B: Input Matrices; C: Output Matrix; g(x): linear map-back function; X_{ij}: i^{th}row,j^{th} column element of matrix X; Δ_1,Δ_2: quantization steps of Matrices A&B; min(X): minimum value of Matrix X; max(X): maximum value of Matrix X; Number of ReRAMs needed = 2n(m+1); Total time taken =(M+1)t_{write}+(N+1)t_{read}; t_{write}: ReRAM programming time; t_{read}:read time; V_x: Output Voltage (V_{out}) x=1,2,3,4; C: capacitance associated with the integrator circuit

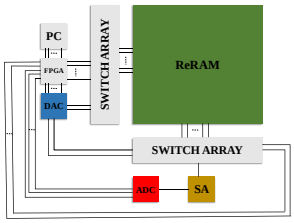


Fig. 11 System level implementation for array testing using Spartan-6 FPGA, ADC/DAC, Integrator (SA) for I/O control and read-out.

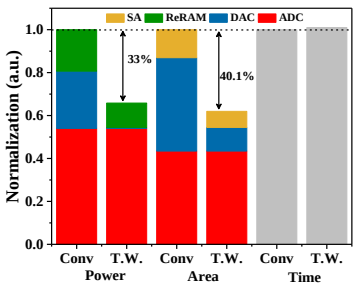


Fig. 12 Comparison of performance parameters. Conv: Conventional, T.W: This Work. Power consumption of system reduced by 33% and area by 40.1%