Four-Layer 3D Vertical RRAM Integrated with FinFET as a Versatile Computing Unit for Brain-Inspired Cognitive Information Processing

Haitong Li[1*], Kai-Shin Li[2#], Chang-Hsien Lin[2], Juo-Luen Hsu[2], Wen-Cheng Chiu[2], Min-Cheng Chen[2], Tsung-Ta Wu[2], Joon Sohn[1], S. Burc Eryilmaz[1], Jia-Min Shieh[2], Wen-Kuan Yeh[2], and H.-S. Philip Wong[1]

[1]Department of Electrical Engineering and SystemX Alliance, Stanford University, Stanford, CA 94305, USA;

[2]National Nano Device Laboratories, NARLabs, Hsinchu, Taiwan; Email: [*]haitongl@stanford.edu; [#]ksli@narlabs.org.tw

Abstract

For the first time, a four-layer HfO_x-based 3D vertical RRAM, the "tallest" one ever reported, is developed and integrated with FinFET selector. Uniform memory performance across four layers is obtained (±0.8V switching, 10^6 endurance, 10^4s@125°C). SPICE simulations show that high drive current of pillar select transistors is required for high-rise 3D RRAM arrays. The four-layer 3D RRAM is a versatile computing unit for (a) brain-inspired computing and (b) in-memory computing. (a) Stochastic RRAM synapses enable robust pattern learning for a 3D neuromorphic visual system. The 3D architecture with dense and balanced neuron-synapse connections provides 55% EDP savings and 74% V_{DD} reduction (enhanced robustness) compared with conventional 2D architecture; (b) in-memory logic such as NAND, NOR, and bit shift, are essential elements for hyper-dimensional computing. Utilizing the unique vertical connection of 3D RRAM cells, these operations are performed with little data movement.

Introduction

Brain-inspired cognitive information processing aims at approaching the efficiency of brain computation where memory and computing are tightly integrated [1]-[3]. In this work, a "tallest"-in-record [4]-[9] four-layer 3D vertical RRAM integrated with FinFET is developed and characterized. The 3D RRAM is a versatile computing unit that not only improves the energy-delay product (EDP) of neuromorphic computing due to dense configuration, but also enables in-memory computing owing to the unique common-pillar (CP) structure.

Device Fabrication and Characterization

Based on an advanced FinFET process platform [10], four-layer vertical RRAM is fabricated after M1 process, consisting of 4-layer 20-nm PVD TiN (BE)/20-nm CVD SiO_2, 5-nm ALD HfO_x, and PVD 10-nm Ti/40-nm TiN (pillar TE) (Fig. 1 and Fig. 2). P-channel FinFET serves as the pillar selector for the 3D RRAM (Fig. 3). Fig. 4 shows the DC switching characteristics of the four-layer 3D RRAM, where the integrated FinFET eliminates the current overshoot during RRAM operations [11]. Statistical distributions of switching parameters show that suitable switching voltage (~±0.8V) and adequate ON/OFF ratio (>10×) are obtained for all four vertical cells (Fig. 5). Endurance tests show stable pulse switching after 10^6 cycles across all four layers (Fig. 6). The measurements are conducted in sequence from L1 up to L4, and no disturbance on adjacent layers occurs during operations. Four-layer cells are also stable after 10^4s retention tests under 125°C (Fig. 7). The device characteristics are then incorporated into full-size 3D array simulations including interconnect RC components using HSPICE [6]. A higher drive current of the select transistor is required to address larger number of vertical layers (Fig. 8). With 400-μA drive current (1-fin FinFET) and 10-kΩ R_{LRS} measured on the fabricated devices, a 32-layer 3D RRAM array can be addressed. Increasing R_{LRS} reduces both the SET current of selected cell and the sneak current from unselected cells, which lowers the I_{ON} requirement of the select transistor.

3D Neuromorphic Architecture

Stochastic learning that embraces the intrinsic variability of RRAM synapses is robust for pattern recognition with the ability of escaping local minima and emulating analog weights [12], [13]. Probabilistic switching of 3D RRAM as stochastic synapses is measured with different suites of pulse amplitude and width to gain a deeper understanding of stochastic learning (Fig. 9). Key observations are: (a) amplitudes lower than actual SET voltage of RRAM can be chosen to achieve a proper switching probability for pattern learning; (b) a shorter pulse requires higher amplitude to reach the same probability due to nonlinear voltage-time relation of RRAM filament evolution (Fig. 9). Such measured behaviors are modeled by the cycle-to-cycle variability of equivalent energy barriers of oxygen vacancy movement, and are incorporated into the variation-aware RRAM compact model we use [14]. The experimentally validated model is then used for the system-level simulations. As an illustration, we study an unsupervised winner-take-all (WTA) visual system consisting of stochastic synapses and leaky integrate-and-fire (LIF) neurons. Conventional networks based on 2D crossbar arrays are unbalanced in structure due to large number of input neurons for receiving the stimuli but small number of output neurons for output classes [12]-[16]. Here, the 3D architecture 'folds' neurons/synapses into balanced (x-y) plane with dense (z direction) connections, and thus reduces interconnect RC effects and avoids long sneak leakage paths of the 2D architecture [17] (Fig. 10). 1000 training images of Gaussian random orientations centered around four dominant angles are fed into the WTA network. After training, four distinct resistance maps are organized. Total energy consumption including contributions from the summing current and energy for programming synapses is obtained (Fig. 11). Shorter pulses require higher amplitudes to reach a certain SET probability. Hence, minimal energy consumption is achieved for a moderate pulse width. Compared with the conventional 2D architecture, the 3D architecture improves EDP by 55% and reduces V_{DD} by 74%, on top of 4× area gain (Fig. 12). The lower array V_{DD} protects un-addressed synapses along the pulse path from being disturbed, which enhances system robustness [17].

In-Memory Computing

Hyper-dimensional computing is error-resilient as information is represented as hyper-dimensional sparse vectors instead of numbers. It has been shown to be effective for cognitive tasks such as language identification [18]. Here, Boolean logic operations are required as basic kernels. The energy efficiency of hyper-dimensional computing can be boosted if in-memory logic is employed to eliminate energy-hungry data movement (aka "von Neumann bottleneck") [19]. Owing to the unique CP structure, essential logic operations are readily realized on the multi-layer 3D RRAM, where the state variable for Boolean logic is the RRAM resistance (R_{LRS}=1, R_{HRS}=0). Two computing modes are available: programming mode (Fig. 13-15) and readout mode (Fig. 16). Any arbitrary multi-stage Boolean expressions can be implemented by programming 3D RRAM cells along a common pillar. Deeper logic stages and multiple-inputs are accommodated by more layer stacks. Specific programming pulse trains are used for basic Boolean operations, such as NAND (Fig. 13) and NOR (Fig. 14). Bit shift from L1 to L4 for hyper-vector permutation can be extremely simple (Fig. 15). For the readout mode, input data are addresses that are presented to a decoder (Fig. 16 inset). The decoder output selects a target pillar (selected by FinFET) that contains the output of the logic function. The previously memorized output data are read out directly without the need to re-program any cell. The endurance of such in-memory computing is of paramount importance. Readout mode is employed for frequently-used logic. 10^{11}-cycle readout operations for NAND/NOR logic evaluations are measured experimentally, limited merely by test time (Fig. 16). The in-memory computing on 3D RRAM is dynamically reconfigurable, from readout mode to programming mode, to allow for new multi-stage logic functions. 10^6 RRAM endurance is estimated to support ~10^5 re-programming for implementing various multi-stage logic functions.

Conclusion

Key achievements: (1) the "tallest" four-layer 3D vertical RRAM integrated with FinFET is developed with uniform performance across four layers; (2) memory-transistor co-design guidelines are provided for high-rise 3D RRAM arrays; (3) 3D architecture with dense neuron-synapse connections improves energy efficiency of neural networks; (4) in-memory computing (NAND, NOR, bit shift) is demonstrated with 10^{11} address-and-read cycles on 3D RRAM. In summary, the balanced array configuration and unique CP structure make 3D RRAM an energy-efficient versatile computing unit, which is not readily achievable by conventional 2D RRAM.

Acknowledgement

This work is supported in part by STARnet SONIC, the NSF Expedition in Computing, Stanford NMTRI and Stanford SystemX Alliance. Device fabrication is performed by NDL facilities and supported by the Ministry of Science and Technology, Taiwan. We thank S. Yu from ASU for his help.

References

[1] H.-S. P. Wong *et al.*, Nature Nanotech., p.191, 2015. [2] M. Prezioso *et al.*, Nature, p.61, 2015. [3] D. Querlioz *et al.*, Proc. IEEE, p.1398, 2015. [4] I. G. Baek *et al.*, IEDM 2011, p. 737. [5] W.C. Chien *et al.*, VLSI 2012, p.153. [6] H.-Y. Chen *et al.*, IEDM 2012, p. 497. [7] E. Cha *et al.*, IEDM 2013, p.268. [8] C.-W. Hsu *et al.*, IEDM 2013, p.264. [9] K.-S. Li *et al.*, VLSI 2014, p.1. [10] M.-C. Chen *et al.*, VLSI 2013, p.218. [11] R. Degraeve *et al.*, VLSI 2012, p. 75. [12] M. Suri *et al.*, IEDM 2012, p.235. [13] D. Garbin *et al.*, IEDM 2014, p.661. [14] S. Yu *et al.*, IEDM 2012, p.239. [15] S. Park *et al.*, IEDM 2012, p.231. [16] G.W. Burr *et al.*, IEDM 2014, p.697. [17] H. Li *et al.*, TED, p.3160, 2015. [18] P. Kanerva, Cognitive Computation, p.139, 2009. [19] S. Borkar *et al.*, Commun. ACM, p.67, 2011.

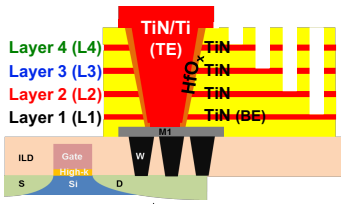

Fig. 1 Schematic of 4-layer 3D RRAM-FinFET & fabrication flow.

- M2 patterning
- 4-layer contact
- HfO$_x$ & TE growth
- Side wall patterning
- 4-layer TiN/SiO$_2$ deposition
- Silicide & M1 contact
- S/D implant & activation
- Gate patterning
- HKMG deposition
- Oxide deposition & etching back
- Fin patterning

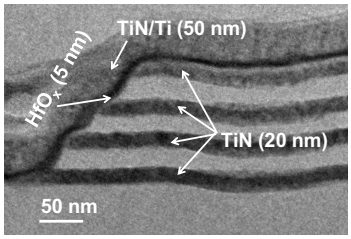

Fig. 2 TEM image of four-layer 3D vertical RRAM. TiN (40 nm)/Ti (10 nm) serves as top electrode, and four TiN layers serve as bottom electrodes. HfO$_x$ across TiN layers is the resistive switching layer.

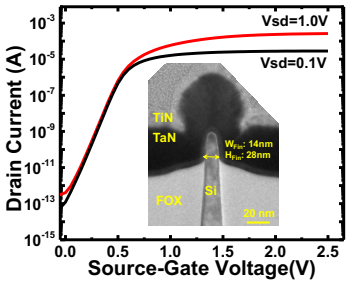

Fig. 3 I_D-V_{SG} characteristics of P-channel FinFET (inset: TEM) with 300-nm gate length and 100-nm width. Good driving capability benefits 3D RRAM operations.

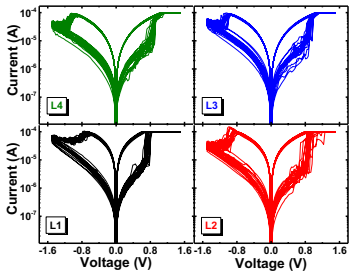

Fig. 4 DC I-V characteristics of four-layer (L1-L4) RRAM cells around the same vertical pillar, with 100-μA compliance current provided by pillar select FinFET.

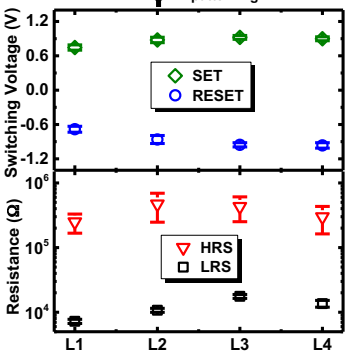

Fig. 5 Measured cycle-to-cycle statistics of resistance and switching voltages of four-layer RRAM cells.

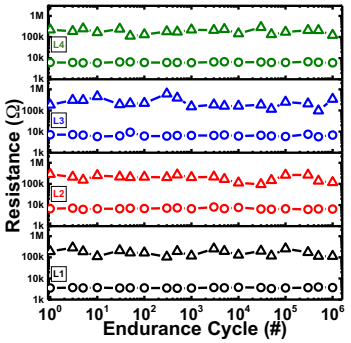

Fig. 6 Measured endurance characteristics. None of four cells show degradation after 10^6 cycles. Besides, consecutive switching is disturb-free on adjacent layers.

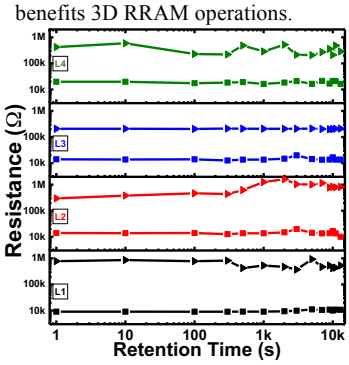

Fig. 7 Measured retention characteristics. Four-layer RRAM cells are stable after 10^4 seconds at 125°C. Read operations are disturb-free on adjacent layers.

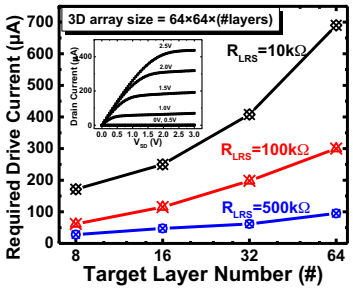

Fig. 8 SPICE-simulated required drive current of bottom transistors for 3D vertical RRAM arrays. Increasing R_{LRS} of RRAM cells mitigates driving capability requirement. Inset shows the measured FinFET I_D-V_{SD} curves.

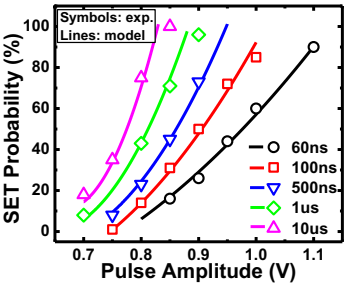

Fig. 9 Measured (symbols) and modeled (lines) probabilistic switching behaviors of RRAM synapses as a function of pulse amplitude and pulse width. Each single probability is determined based on 100 SET/RESET pulse cycles (RESET: -1.5 V pulses).

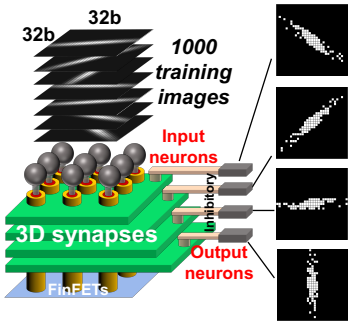

Fig. 10 A 3D neuromorphic visual system based on a 32×32×4 3D array. A winner-take-all (WTA) network with stochastic learning rule is simulated for orientation detection. Four-layer synapses are self-organized into distinct R-maps.

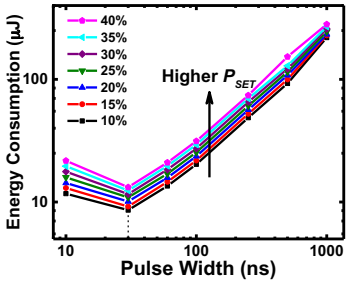

Fig. 11 Simulated total energy consumption for training the WTA network as a function of pulse width and synapse SET probability (P_{SET}). Shorter pulses require higher amplitude for the certain P_{SET}, which leads to optimal energy point found at 30 ns pulse width.

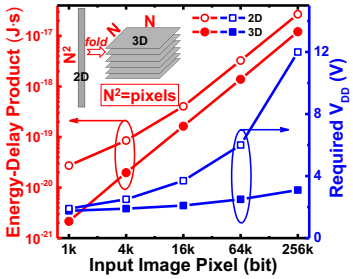

Fig. 12 SPICE-simulated EDP and required V_{DD} to program worst-case-located synapses (15% P_{SET}) in 2D and 3D WTA network. 3D architecture leads to 55% EDP savings and 74% V_{DD} reduction (better reliability) on 256-kb input images compared with 2D design.

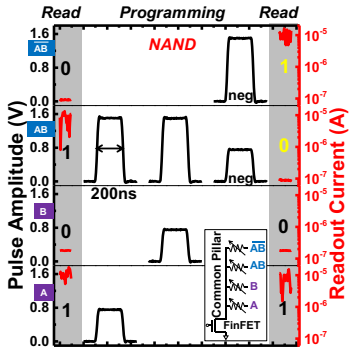

Fig.13 Experimental demonstration of in-memory NAND logic on the CP structure (inset). Waveforms show the applied pulse train (black) to perform computation and the initial/final readout states (red) of input/output RRAM cells. NAND/AND are vertically realized in ~8F^2 area.

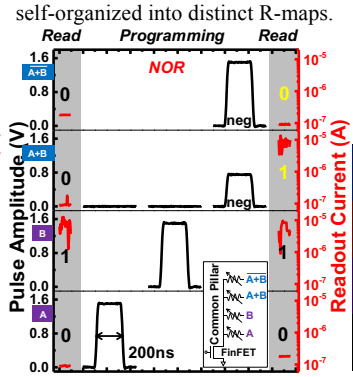

Fig.14 Experimental demonstration of in-memory NOR logic on the CP structure (inset). Waveforms show the applied pulse train (black) to perform computation and the initial/final readout states (red) of input/output RRAM cells. NOR/OR are vertically realized in ~8F^2 area.

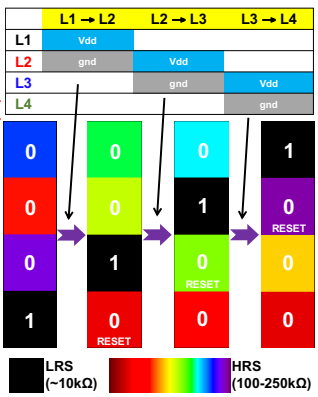

Fig. 15 Measured inter-layer bit shift illustrated by resistance evolution (upper shows timing diagram). LRS ('1') is in black and HRS ('0') is in rainbow color scale. Bit '1' is being shifted vertically from L1 to L4.

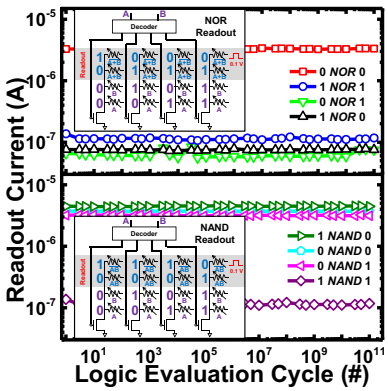

Fig. 16 Measured in-memory computing endurance with no error bit after 10^{11} readout cycles of memorized functions. Inset shows circuit implementation of logic evaluations via address-and-read, without the need to re-program RRAM cells (switching endurance requirement is alleviated).