Design and Optimization Methodology for 3D RRAM Arrays

Yexin Deng1†, Hong-Yu Chen2, Bin Gao1, Shimeng Yu1, Shih-Chieh Wu1, 2, 3, Liang Zhao2, Bing Chen1, Zizhen Jiang2, Xiaoyan Liu1, Tuo-Hung Hou3, Yoshiho Nishi2, Jinfeng Kang1*, and H.-S. Philip Wong2#

1 Institute of Microelectronics, Peking University, Beijing 100871, China;
2 Department of Electrical Engineering and Center for Integrated Systems, Stanford University, Stanford, CA 94305, USA;
3 Department of Electronics Engineering and Institute of Electronics, National Chiao Tung University, Hsinchu, Taiwan;

Email: †deng58@purdue.edu, *kangjf@pku.edu.cn, #hspwong@stanford.edu

Abstract - 3D RRAM arrays are studied at the device- and architecture- levels. The memory cell performance for a horizontal cross-point is shown experimentally to be essentially comparable to vertical pillar-around geometry. Array performances (read/write, energy, and speed) of different 3D architectures are investigated by SPICE simulation, showing horizontal stacked RRAM is superior but suffers from higher bit cost. Adopting a bi-layer pillar electrode structure is demonstrated to enlarge the array size in 3D vertical RRAM. Design guidelines are proposed for the 3D VRRAM: it shows that increasing the number of stacks of VRRAM while keeping the total bits the same, as well as scaling of feature size (F), are critical for reducing RC delay and energy consumption.

I. Introduction

3D RRAM arrays are promising for future mass storage application [1]. Recently, the horizontally stacked 3D RRAM (HRRAM) architecture and several novel vertical 3D RRAM (VRRAM) architectures have been experimentally demonstrated [2-4]. In particular, we have proposed and demonstrated a cost-effective 3D VRRAM array together with the array operation as well as analyzed the scaling limits [5]. The limitation on achievable array size is mainly due to interconnect resistance which degrade the read/write margin due to I×R drop in the interconnect [6]. However, the broader performance metrics (speed, energy) among different 3D RRAM arrays have not been systematically studied, and the design methodologies of 3D RRAM still needs to be developed. In this paper, horizontally stacked 3D RRAM (HRRAM) [2] and two variants of bit-cost scalable vertical 3D RRAM (VRRAM_1 [3], VRRAM_2 [4]) arrays are studied from both the device and the architecture point of view. (Figs. 1-3).

II. Device-Level Study:

The only difference among the three array designs at the device level is the geometry of the memory cell. In particular, to maximize device density, the diameter of the pillar should be minimized in VRRAM_2. Thus we propose a core-shell pillar structure (Fig. 4) that has a low-resistance metal core and a TiN shell for RRAM functionality. It can enlarge the array size substantially, especially when F scales to sub-20 nm region (Fig. 5). To study how thin the TiN can be while maintaining RRAM functionality, we prepared three TiN/HfOx/Pt RRAM samples with the same cross-point areas (~1μm²): (S1) is an improved version of the vertical cell with tm scaled to 5 nm [5] (Fig. 6), (S2) is a horizontal cell that serves as a control sample and (S3) is a horizontal cell with bi-layer top electrode (TE) W/TiN (~ 3 nm) (Fig. 7). S3 is a proof-of-concept test structure for the proposed core-shell structure in [5]. High resolution EELS analysis (Fig. 8) confirms the thickness of ultra-thin TiN is ~3 nm. S1, S2 and S3 show similar resistive switching behaviors such as I-V characteristics (Fig. 9), resistance distribution (Fig. 10), and switching voltage distributions (Fig. 11). Device-to-device variation for the bi-layer TE sample (S3) is reasonably controlled in terms of the switching voltage and HRS/LRS distribution from 10 different cells (Figs. 12&13). We conclude that (1) the impact on device performance for horizontal and vertical pillar-around cell geometries is insignificant at the device level, and (2) the TiN layer functions well as oxygen reservoir for RRAM functionality while W capping layer functions successfully as a contact/interconnect, suggesting that the core-shell pillar structure should be realizable in order to reduce the interconnect resistance [5]. Other conductive metals that do not interact with TiN may also be considered for the core metal material in the future.

III. Architecture-Level Study

a) Experimental Measurement: The difference among three arrays at the architecture level is the wire routing of memory cells. The interconnect resistance and capacitance play important roles in array performance, and the effects are experimentally observed in VRRAM_2. A larger voltage/longer time is required to switch the cell farther from the point on which voltage is applied (Fig. 14). This implies that the applied voltage drop on interconnect can be significant, particularly when tm is scaled in VRRAM_2. Furthermore, the parasitic components increase the
switching time and limit the array speed in terms of RC delay. The delay time for various plane electrode distances is measured by reading out the transient current. The test structure and measurement setup are illustrated (Fig.15). The estimated order of ~ ns RC delay was observed when the voltage is applied far from the device (Fig. 16).

b) Comparison of Architectures Performances: To systematically study the three array designs at the architecture level, full-size SPICE circuit models for three 3D architectures are constructed including (1) interconnect resistance/capacitance of different geometry and (2) resistance/capacitance of RRAM cells. Write access voltage, readout margin, RC delay time, and energy consumption per write operation as a function of total number of bits for the HRRAM, VRRAM_1 and VRRAM_2 arrays are simulated respectively under worst-case conditions (Figs. 17-20). The criterion to estimate the total number of array bits during write operation (Fig. 17) is set at 2/3 Vdd (Vdd=3V) for the worst-case cell. We find that (1) HRRAM can achieve larger array size and shorter RC delay while maintaining low energy consumption, and (2) VRRAM_2 shows larger readout margin during read operation. However, it is worth noting that HRRAM may demand a higher fabrication cost due to the stack-related lithography process [1].

c) Optimization Guideline for VRRAM_2: It is useful to explore if the bit-cost effective VRRAM_2 can achieve the performance comparable to HRRAM. Design guidelines are explored by optimizing the dimension parameters and device characteristics (Fig. 4). Since read/write operations and their relation to improve total bits and bit density were discussed in [5], here we focus on the RC delay and energy consumption per write operation of the array including the wires but excluding the readout circuit to keep our results general. Both static and dynamic energy are considered and the pulse width is assumed to be 5 ns with 3V amplitude. First, the RC delay of a 1Mb array is found to be larger than the fastest experimentally reported switching time (300 ps) of RRAM [6] (Fig. 21). Then, increasing the number of stacks from 4 to 64, while keeping the total number of bits in the array the same, will reduce the delay by 7× and reduce energy by 5× (Figs. 22&27), respectively. This is because increasing the number of stacks will reduce the plane area, so the parasitic C and the leakage paths from the selected plane to neighboring half-selected planes will decrease. The delay is also reduced by a half when F scales from 30 nm to 20 nm (Fig. 23) and the trend is also shown in the sub-20 nm region. Scaling down the thickness of metal plane electrode (tm) and isolation layer (ti) can worsen the delay time (Fig. 24). The static energy consumption is found to dominate the total energy, which implies that the leakage energy consumption is significant (Fig. 26). By increasing Ron (from 100kΩ to 1MΩ), the energy consumption can be reduced by ~4× for a 1Mb array (Fig. 28), but this will increase the delay time by ~2× for an on/off ratio=100 (Fig. 25). Moreover, by constraining the leakage paths, the nonlinearity of Ron allows the energy consumption to be reduced down to ~20 pJ even for a 1Mb array (Fig. 29). Finally, the effect of the vertical transistor serving as bit-line selector is analyzed. For the case of Ion=100μA with Ron=100kΩ, the array size cannot achieve even a 1kb array (Fig. 30), so a larger Ron is necessary. Based on the experimental results of vertical transistor reported in [7], write access voltage as F scales down is shown in Fig. 31 with an increased Ron (1MΩ or even 10MΩ). As F scales down to 20 nm, the total number of bits achievable is only 50kb. Therefore, to enlarge array size up to 1Mb or even 10Mb, Ron=10MΩ is required. Nevertheless, this will significantly increase the delay time. Based on all analyses above, the design tradeoffs are summarized in Table 1.

IV. Conclusion
Key achievements: 1) Geometry effect of HRRAM and VRRAM does not impose significant negative influence on the device characteristics for the pillar-around type memory cell. 2) The W/TiN structure with ~3nm TiN is demonstrated as a proof of core-shell pillar concept and presents reliable characteristics. 3) Performance of HRRAM shows superior performance but suffers from higher bit cost compared to VRRAM architectures. 4) Based on the criterion of RC delay, energy consumption and transistor effect, design guidelines are presented for 3D VRRAM.

Acknowledgement
This work is supported in part by 973 Program (2011CBA00600), NSFC (61334007), and the member companies of the Stanford Non-Volatile Memory Technology Research Initiative (NMTRI), STARnet SONIC. This work was done in part at the Stanford Nanofabrication Facility (SNF), a member of the NSF-supported National Nanotechnology Infrastructure Network (NNIN).

Reference
Fig. 1 Schematic of 3D horizontal stacking RRAM array (HRRAM) [2]. The memory cell is between the cross point of word line and bit line.

Fig. 2 Schematic of 3D vertical RRAM (VRRAM_1) array proposed in [3]. The memory cell is between the horizontal word line and the vertical pillar.

Fig. 3 Schematic of 3D vertical RRAM (VRRAM_2) array proposed and demonstrated by us in [4, 5]. The memory cell is between the plane electrode and the vertical pillar.

Fig. 4 Schematic of metal core-shell pillar structure and the dimension parameters of VRRAM_2 array.

Fig. 5 Comparison of array size achieved by Cu-W-TiN core-shell pillar and TiN pillar by the SPICE simulation. As F scales down, the array size can be increased by 400×/100× by using the Cu-W-TiN core-shell pillar.

Fig. 6 High resolution TEM image of scaled 5 nm metal plane electrode VRRAM_2 sample. The sidewall is improved compared to [5].

Fig. 7 HRRAM samples are fabricated (left) and high resolution TEM image of core-shell W/TiN bi-layer structure (right). The thickness of TiN is estimated to be ~3 nm. The cap W serves as a contact/interconnect to reduce the interconnect resistance.

Fig. 8 Electron energy loss spectroscopy (EELS) composition profile through the device. The nitrogen (N) signal confirms the thickness of TiN layer is estimated to be ~3 nm.

Fig. 9 Typical DC I-V switching characteristics. Similar behavior among the V-RRAM sample, H-RRAM control sample and bi-layer H-RRAM sample.

Fig. 10 Resistance distribution obtained from 100 DC cycles. The mean of HRS/LRS window>10. The differences among the three samples are not significant.

Fig. 11 Set and reset voltage distribution are similar among the V-RRAM sample, H-RRAM control sample and bi-layer H-RRAM sample.

Fig. 12 Set and reset voltage distribution for 10 different bi-layer HRRAM sample cells. The cycle to cycle (error bar) and device-to-device (x-axis) variation is reasonably controlled.

Fig. 13 HRS (high resistance state) and LRS (low resistance state) resistance distribution for 10 different bi-layer H-RRAM cells. Good device-to-device uniformity is observed. HRS shows larger variation compared to LRS.

Fig. 14 Set and reset voltage as a function of pulse width collected from metal plane BE1 (near by the pillar TE) and distant BE2. Greater voltages required for the distant electrode suggests the voltage drop on the metal electrode plane is obvious.

Fig. 15 Measurement setup to estimate the access delay time on the fabricated 2-layer vertical RRAM array. Oscilloscope is connected in series with RRAM to measure the transient current through the device. Different distance between voltage applied point and the device is used.

Fig. 16 Measured transient current on the selected device for different distances between the TE of the device and various voltage applied points on the plane bottom electrodes (BE). The delay time of ~ns is observed for different distances.
(2/3)V_{dd} (2V) is assumed to be the lowest switching voltage for the purpose of estimating the array size. Larger array can be achieved by HRRAM.

Fig. 21 Access delay time as a function of array size. 300ps is the shortest switching time of RRAM [5]. A 10Mb array takes 2250um² array area and ~4ns delay time.

Fig. 26 Write energy consumption per operation as a function of total bits. The static energy consumption dominates the total energy consumption, suggesting that the leakage current power is significant in the VRRAM_2 array.

Fig. 27 Write energy as a function of arrays with different number of stacks. 4 stacks array shows 5 times larger energy consumption than 64 stacks array with the same total bits. This can be attributed to larger selected-plane to half-selected plane leakages in 4 stacks array.

Fig. 28 Write energy as a function of total bits. By increasing R_{on} (from 100kΩ to 10MΩ), the energy consumption can be reduced (~4X for a 1Mb array). However, this will increase the delay time (~2X) as shown in Fig. 25 if assuming on/off ratio=100.

Table 1 Summary of Design Tradeoffs

<table>
<thead>
<tr>
<th>Impact of increasing different factors on the performance of the 3D VRRAM. The number of the arrows represents the degree of the impact of the factor. (*The total number of bits is kept the same)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
</tr>
<tr>
<td>bits</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>T_{m}</td>
</tr>
<tr>
<td>R_{on}</td>
</tr>
<tr>
<td>Stacks</td>
</tr>
</tbody>
</table>

Fig. 31 Write access voltage with larger R_{on} (1MΩ and 10MΩ), when F scales from 30nm to 20nm. The I_{on} scaling of vertical transistor with F is from [7]. When F scales to 20nm, only a small array can be achieved.