Selector Requirements for Tera-Bit Ultra-High-Density 3D Vertical RRAM

Zizhen Jiang[1*], Shengjun Qin[1], Haitong Li[1],
Shosuke Fujii[1, 2], Dongjin Lee[1, 3], Simon Wong[1], and H.-S. Philip Wong[1#]
[1]Dept. of Electrical Engineering and Stanford SystemX Alliance, Stanford University, Stanford, CA 94305, U.S.A;
[2]Device Technology R&D Center, Toshiba Memory Corporation, Yokkaichi, 512-8550, Japan;
[3]Seminconductor R&D Center, 1, Samsungjeonja-ro, Hwaseong-si, Gyeonggi-do 18848, South Korea;
E-mail: *jiangzz@stanford.edu, #hspwong@stanford.edu.

Abstract: Selector requirements for tera-bit class, ultra-high-density 3D vertical resistive random access memory (VRRAM) are presented, including practical design considerations such as array efficiency (AE), pillar driver transistors (pillar drivers), and wire/metal plane resistances. We design a novel chip architecture that is different from 3D NAND: (a) separated, square and large wordplane (WP) connected by global wordplane connections (WPC) within a block to minimize influence of leakage currents, (b) compact staircase. An accurate, computationally efficient resistor network is developed to model the parasitic resistances of the architecture. Through the resistor network simulations, selector requirements for 3D VRRAM are examined. To achieve tera-bit class 3D VRRAM with density higher than the most advanced 3D NAND flash (> 4.3 Gb/mm^2), selector nonlinearity (NL) $\geq 10^2$ is required.

I. Introduction: Resistive random access memory (RRAM) offers bit-alterability, direct over-write, fast programming speed (< 10 ns), and low energy consumption (< pJ) for high-density, on-chip data storage [1]. To compete with the most advanced 3D NAND flash (768 Gb and 4.3 Gb/mm^2) [2-4], suitable chip architectures for 3D VRRAM [5-8] need to be identified. Selectors are required to reduce leakage currents and increase array size, AE, and bit-density [9-11]. This paper presents an architecture for tera-bit class ultra-high-density 3D VRRAM and analyzes the selector requirements.

II. Chip Architecture: The selector requirements **must** be analyzed in the context of the chip architecture. Fig. 1 shows the floor plan of a 3D VRRAM that achieves 1 Tbit and 4.6 Gb/mm^2. There are 64 planes, each paired with a row decoder. 512 blocks are vertically aligned in each plane. Fig. 2 shows that each block has 64 layers and is further divided into 8 arrays, and each array has 4 M RRAM cells. The bitline (BL) controller and additional peripheral circuits are at the bottom of the chip. Row decoders and the BL controller guarantee random bit access. Each row decoder can arbitrarily select one wordline (WL) and one WPC. Each WPC is connected to a specific layer of WPs in a block. WPC are placed on top of the arrays, and are wide and tall with low resistivity (0.025 Ω•μm [12]). WLs and WPCs run the entire horizontal length of each plane. BLs run the entire vertical length of each plane. The BL controller and the peripheral circuits can sense one or multiple BLs. Fig. 3-4 show the 3D and top view of an array within a block. FinFETs underneath the array are used as pillar drivers. Hexagonal pillar pattern maximizes the areal density. In 3D NAND, the word line can select only four rows of cells in the same layer. In this 3D RRAM architecture, the WP can select a large array of cells. Using a larger array size or more arrays in each block can reduce the number of planes and the number of decoders, and therefore maximizes AE. Additionally, square arrays (#WL = #BL) are preferred to lower the parasitic WP resistance for accessing the worst-case selected cell. Because the WP is large and square (in contrast to 3D NAND), a compact staircase [13] (Fig. 3) can be used to connect the WP to the corresponding WPC, which further increases AE.

III. Pillar Driver: FinFETs [14-17] are chosen for the pillar drivers due to their higher saturation current and larger pillar areal density (Fig. 5). The minimum unit cell area is determined by the gate pitch and metal pitch (since VRRAM pillar can be smaller than NAND); it can be reduced to 12 F^2 using a single-fin pillar driver using 7 nm FinFET design rules [17], potentially enabling densities of 10 Gb/mm^2. To simplify the following analysis, the more relaxed two-fin layout (Fig. 6) is applied. With the driver saturation current < 50 μA, Low Resistance State (LRS) > V_{DD}/50 μA is required.

IV. Modeling: A resistor network is built to simulate the worst-case scenarios (Fig. 7). The network captures all leakage paths of the arrays within a block. The WP resistances are obtained by 2D field solver (Sentaurus) simulations (Fig. 8). The selector and the RRAM (1S1R) are sandwiched between the WP and the metal pillar (Fig. 9). The 1S1R is simplified as a voltage-dependent resistor. LRS is the equivalent resistance of 1S1R at V_{DD}, while Unselected Resistance State (URS) is the equivalent resistance at 0.5V_{DD} (Fig. 10) [18]. NL of 1S1R is defined as the ratio between URS and LRS. To simulate (Table I) a mega-bit array in a computationally efficient manner, we simplified the 64-layer network to 2-layer equivalent network (Fig. 11). The network is applicable to both write and read programming with < 2% error verified for sub-Mb arrays (Fig. 12).

V. Selector Requirements: Using the resistor network, we explore the number of WLs and BLs in each array and the numbers of arrays, blocks, and planes that meet the storage capacity and the bit-density requirements. The bit-density calculation takes into account the areas of the decoders, the controller, and peripheral circuits, which are estimated to be the same as those of 3D NANDs [4]. To compete with the density of 3D NAND flash (4.3 Gb/mm^2) while achieving adequate read/write margins (criteria in Table I) requires LRS = 1 MΩ and NL $\geq 10^2$ (Fig. 13), yielding 4.6 Gb/mm^2 for single-bit per cell storage. This high density is attained because of smaller pillar dimensions and more compact staircase of 3D VRRAM, as compared to 3D NAND designs. Higher NL and multi-level programming of RRAM devices can provide even higher density (6.3 Gb/mm^2). Segmented read (not reading all BLs simultaneously, trading off bandwidth/latency) and bias optimization can loosen the 1S1R requirements. The design specifications, the resistances and the resistivities used in the simulation are summarized in Table II and Table III.

VI. Conclusion: We present 1) a 3D VRRAM architecture for tera-bit class storage competitive with 3D NAND; 2) a 3D VRRAM-specific architecture that significantly loosens 1S1R requirements for high-density storage; 3) an accurate and computationally efficient resistor network. Single-bit/cell, tera-bit class high-density (> 4.3 Gb/mm^2) 3D VRRAM is achievable for LRS = 1 MΩ and selector NL $\geq 10^2$.

Acknowledgement: This work is supported in part by Stanford NMTRI and the Samsung Electronics' University R&D program.
Reference: [1] Stanford Memory Trends; [2] Tanaka, ISSCC '16; [3] Kim, ISSCC '17; [4] Yamashita ISSCC, '17; [5] Baek, IEDM '11; [6] Chien, VLSI '12; [7] Chen, IEDM '12; [8] Liu, ISSCC '13; [9] Kim, VLSI '12; [10] Lee, VLSI '12; [11] Cha, IEDM '13; [12] Lee, IEDM '16; [13] Chen, IEDM '12; [14] Stanford CMOS Technology Scaling Trend; [15] Yang, EDL '08; [16] Wu, VLSI '17; [17] ASU, ASAP, 7 nm PDK; [18] Govoreanu, ICICDT '15.

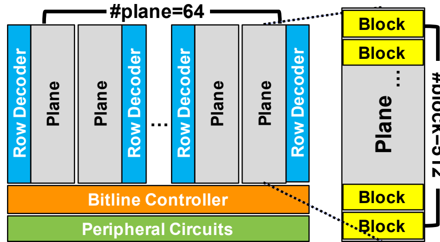

Fig. 1. Floor plan of a 3D VRRAM chip of 1 Tbit, 4.7 Gb/mm^2.

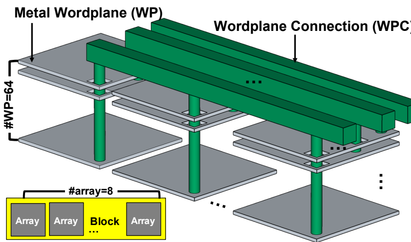

Fig. 2. Schematic of 3D arrays in a block. Low resistance WPC connects with all arrays of a layer.

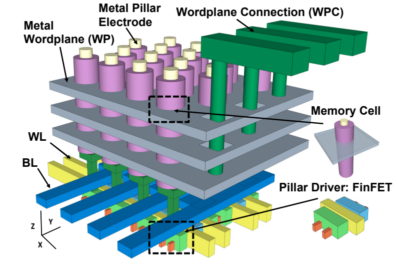

Fig. 3. View of a 3D VRRAM array.

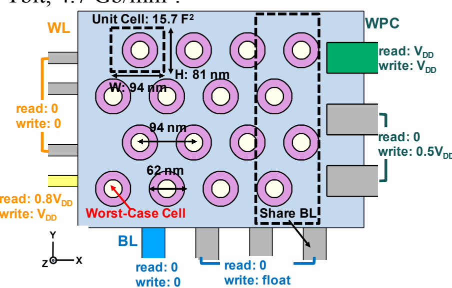

Fig. 4. Top view of 3D VRRAM array.

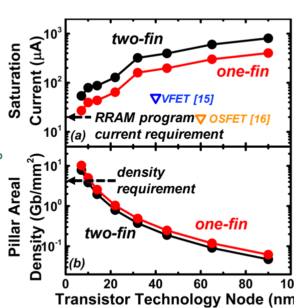

Fig. 5. Pillar driver saturation current and areal density for various transistor technology nodes. Data points for node > 22 nm indicate the equivalent transistor size scaled up from 7 nm FinFET [17]. Vertical FET (VFETs) [15] or BEOL-compatible oxide semiconductor FET (OSFETs, e.g. IGZO FET) [16] can be used if they can provide enough current for RRAM programming at the same areal density.

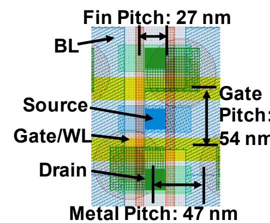

Fig. 6. Layout of pillars with two-fin transistors.

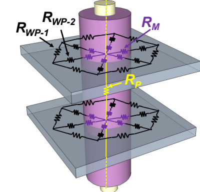

Fig. 7. Schematic of a sub-circuit module that captures the distributed resistances.

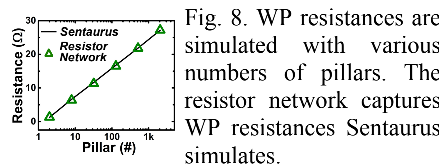

Fig. 8. WP resistances are simulated with various numbers of pillars. The resistor network captures WP resistances Sentaurus simulates.

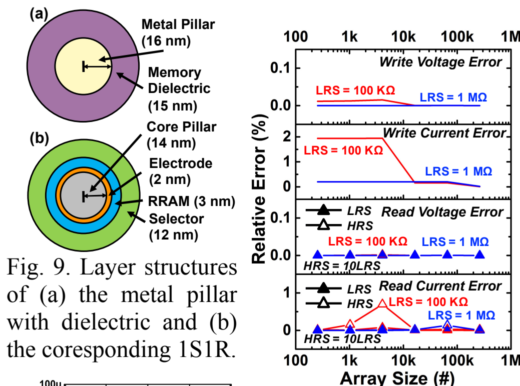

Fig. 9. Layer structures of (a) the metal pillar with dielectric and (b) the coresponding 1S1R.

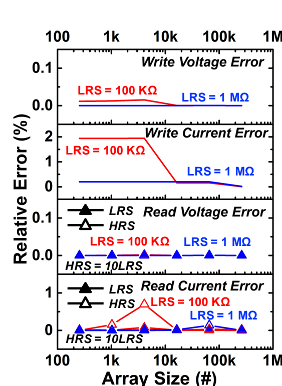

Fig. 12. Relative errors of full and reduced resistor networks on the access voltage of the worst-case cell and the current through the corresponding BL respectively. The relative errors are below 2%.

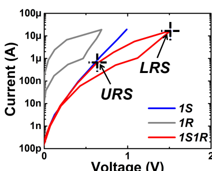

Fig. 10. Characteristics of typical selector (1S), RRAM (1R), and 1S1R devices.

Table I. Bias conditions, corresponding 1S1R resistances and write/read criteria.

Parameter	Details	Unit	Write	Read
V_{WPs}	selected WP applied voltage	V	V_{DD}	$0.8V_{DD}$
V_{WPu}	unselected WP applied voltage	V	$0.5V_{DD}$	$0.5V_{DD}$
V_{WLs}	selected WL applied voltage	V	V_{DD}	V_{DD}
V_{WLu}	unselected WL applied voltage	V	0	0
V_{BLs}	selected BL applied voltage	V	0	0
V_{BLu}	unselected BL applied voltage	V	float	float
R_{Ms}	selected 1S1R equivalent resistance	Ω	LRS	Read LRS: LRS Read HRS: HRS
R_{Mu}	unselected 1S1R equivalent resistance	Ω	URS	URS

Write criteria: 1) $V_{WC} > 0.9V_{DD}$; 2) Pillar driver can provide enough current.
Read criteria: 1) $V_{WC} > 0.65V_{DD}$; 2) $I_{LRS} > 100$ nA; 3) $I_{LRS} > 5I_{HRS}$.
Applied voltages for the worst-case cell analysis is shown in Fig. 4. V_{WC} is the voltage drop on the worst-case cell and I_{LRS} and I_{HRS} are the currents through the pillar driver when reading the worst-case cell at LRS and HRS respectively.

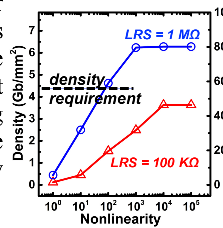

Fig. 13. LRS = 1 MΩ and NL ≥ 10^2 are required for tera-bit class 3D VRRAM with > 4.3 Gb/mm^2. The maximum density is limited by LRS and the areas of the row decoders, BL controller and the additional peripheral circuits.

Table II. Design specifications for different NL requirements of 1Tbit 3DVRRAM (LRS=1 MΩ and #WP=64).

NL	AE (%)	Density (Gb/mm^2)	#plane	#block	#array	#WL	#BL
10^2	59	4.6	64	512	8	256	256
10^3	80	6.2	16	256	16	512	512
10^4	80	6.3	8	128	16	1024	1024
10^5	80	6.3	8	128	16	1024	1024

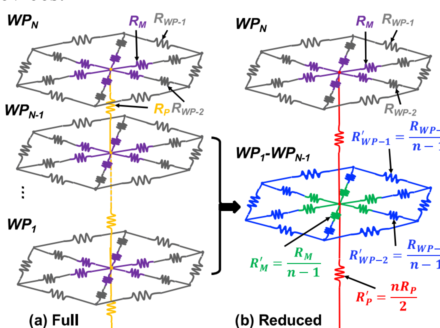

Fig. 11. Schematic of (a) full and (b) reduced resistor networks. The unselected WPs (WP_1-WP_{N-1}) can be combined as one equivalent WP. The resistances of the reduced network are provided in the schematic.

Table III. Resistance (R) and resistivity (ρ) used in the simulation. The definition can be found in Fig. 7 (WP-1, WP-2, and P) and Fig. 4 (BL and WL).

	R (Ω)	ρ (Ω·μm)
WP-1	54.4	0.2
WP-2	27.2	0.2
P	0.57	0.042
BL	1.3	0.035
WL	7.5	0.064