Next-Generation Ultrahigh-Density 3-D Vertical Resistive Switching Memory (VRSM)—Part II: Design Guidelines for Device, Array, and Architecture

Zizhen Jiang, Shengjun Qin, Haitong Li, Shosuke Fujii, Dongjin Lee,
S. Simon Wong, and H.-S. Philip Wong

***Abstract*—Using the reduced resistor network developed in Part I of this two-part article, we present practical design guidelines from device to architecture levels to achieve ultrahigh-density 3-D vertical resistive switching memory (VRSM). We first design both hexagon and comb arrays using 7-nm FinFET as pillar driving transistors (pillar drivers). Small-footprint pillar drivers are necessary for a high pillar areal density competitive to 3-D NAND. We then organize the arrays into an architecture using the compact staircase and highly conductive wordplane connection (WPC) to maximize array efficiency and chip density. We investigate the memory and selector requirements, tolerance of parasitic resistances, latency, and energy consumption for both hexagon and comb architectures. The results indicate that the hexagon array with large low-resistance state (LRS) and nonlinearity (NL) is required for ultradense 3-D VRSM. Compared to the comb array, the hexagon array benefits from a continuous WP pattern and yields a better tolerance of parasitic resistances and a smaller latency. The energy consumptions of both architectures are similar. Compared to the most advanced 3-D NAND, 3-D VRSM has higher chip density and shows better potential for future ultradense storage.**

***Index Terms*—3-D, architecture, array, conductive bridge random access memory (CBRAM), nonlinearity (NL), phase change memory (PCM), resistive switching memory (RSM), resistive random access memory (RRAM), selector, ultrahigh density.**

I. Introduction

RESISTIVE switching memories (RSMs), resistive random access memory (RRAM), conductive bridge random access memory (CBRAM), and phase change memory (PCM) are competitive candidates for future high-density storage due to their fast-speed (∼5 ns), low energy consumption (<pJ), CMOS compatibility, bit alterability, and direct overwrite [1]. Previous works demonstrated the concept [2]–[6], functionality [7], [8], and scaling [7], [9], [10] of 3-D vertical RSMs (VRSMs). Baek *et al.* [2], Chien *et al.* [3], and Chen *et al.* [4] demonstrated the integration of TaO_X-, WO_X-, and HfO_X-based 3-D vertical RRAM (VRRAM), respectively. Later, 3-D vertical CBRAM [5] and 3-D vertical PCM [6] were also reported. Among those previous works, two typical array structures were reported, each with tradeoffs. The continuous metal plane array uses a one-step etch process to pattern the metal wordplanes (WPs) and the memory holes (MHs) at the same time [4], while the chopped metal plane array (comb array) requires an additional etch process to pattern the MHs [2], [3]. The comb array has smaller capacitance and double the number of cells per pillar with the patterned metal planes [2], [3]. Deng *et al.* [7] and Zhang *et al.* [8] demonstrated the functionality of both kinds of 3-D arrays. To increase the array density, researchers have explored thinning down the thickness of the electrode [7], metal plane [9], isolation oxide [9], and memory dielectric [10]. However, the remaining challenge is how to achieve a chip with bit density larger than the current 3-D NAND. The density of 3-D 64- and 128-layer NAND are already 4.3 and 7.8 Gb/mm^2, respectively[11]–[14]. Toward this goal, we previously presented the design for a 1-Tb, 6.3-Gb/mm^2 3-D 64-layer VRRAM using a continuous metal plane array (hexagon array of pillars) [15]. In this two-part article, we present an accurate and computationally efficient reduced resistor network, analyze the chip architecture using both hexagon and comb arrays (see Fig. 1), and provide design guidelines for future ultrahigh-density 3-D VRSM with different WP layers, targeting densities higher than 3-D NAND.

In Part I of this article [16], we defined the nonlinearity (NL) for one selector one resistive switching memory (1S1R) and described our simulation platform. We developed a full

Manuscript received August 14, 2019; revised October 3, 2019; accepted October 19, 2019. Date of publication November 18, 2019; date of current version November 27, 2019. This work was supported in part by the member companies of the Stanford Non-Volatile Memory Technology Research Initiative (NMTRI) and in part by the Samsung Electronics' University R&D Program. The review of this article was arranged by Editor J. Kang. *(Corresponding author: Zizhen Jiang.)*

Z. Jiang, S. Qin, H. Li, S. S. Wong, and H.-S. P. Wong are with the Department of Electrical Engineering and Stanford SystemX Alliance, Stanford University, Stanford, CA 94305 USA (e-mail: jiangzz@stanford.edu; hspwong@stanford.edu).

S. Fujii is with the Department of Electrical Engineering and Stanford SystemX Alliance, Stanford University, Stanford, CA 94305 USA, and also with Kioxia Corporation, Yokkaichi 512-8550, Japan.

D. Lee is with the Department of Electrical Engineering and Stanford SystemX Alliance, Stanford University, Stanford, CA 94305 USA, and also with the Samsung Semiconductor R&D Center, Hwaseong 18848, South Korea.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2019.2950595

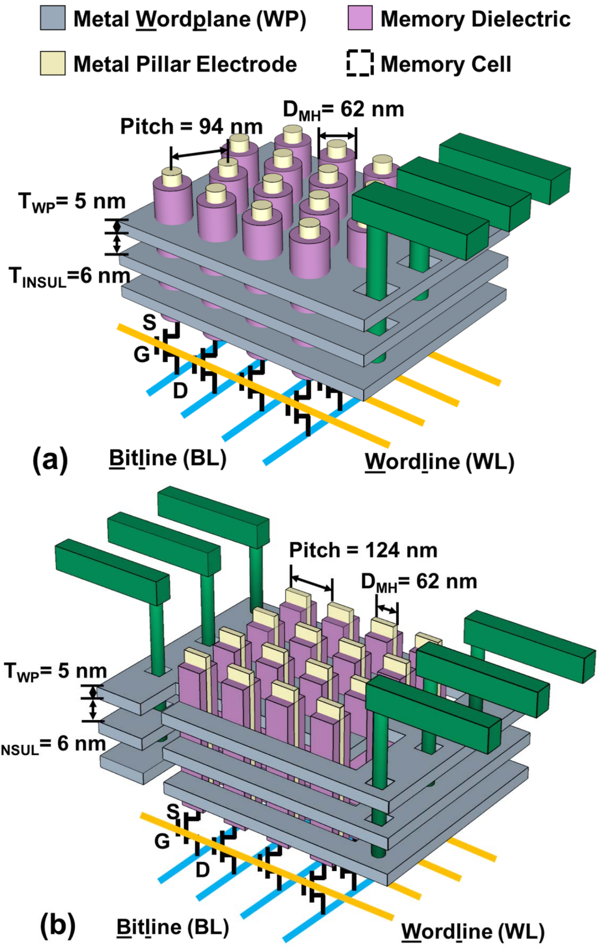

Fig. 1. Schematic of 3-D VRSM array. (a) Continuous memory plane array (hexagon array). (b) Chopped memory plane array (comb array).

resistor network to simulate the 1R and 1S1R hexagon array accurately compared with the simulation results of a 2-D field solver (Sentaurus). To improve computational efficiency, we simplified the resistor network into a reduced resistor network, which significantly reduced the simulation time and memory usage. The relative error of the reduced resistor network was less than 2%. We also presented a comparison of write/read margins between the 1R and 1S1R arrays, indicating that the array size can be increased to megabit scale with $\mathrm{NL} = 10^3$. In Part II of this article, we investigate the device requirements for designs of array and architecture that achieve high cell density. We present the layouts of both hexagon and comb arrays with shared-source pillar drivers and devised an ultradense 3-D VRSM architecture using compact staircases and WP connections (WPCs). We show that the hexagon array with large low-resistance state (LRS) and adequate NL is preferred for high-density storage. Compared to 3-D NAND, 3-D VRSM has good potential to achieve denser storage.

This article is organized as follows. Section II presents the practical designs of array and architecture. Section II-A describes the layout of the hexagon and the comb arrays using 7-nm FinFET design rules. Section II-B lays out the floor plan of the ultrahigh-density chip architecture. Here, we provide an example of 3-D VRSM architecture that is denser than 3-D NAND. Section III illustrates the corresponding device requirements. Section IV evaluates the maximum acceptable parasitic resistances of WP and pillar for the dense 3-D VRSM architecture. Section IV also provides further discussion and comparison between hexagon and comb 3-D VRSM architectures (see Section IV-A) and between 3-D VRSM and 3-D NAND (see Section IV-B).

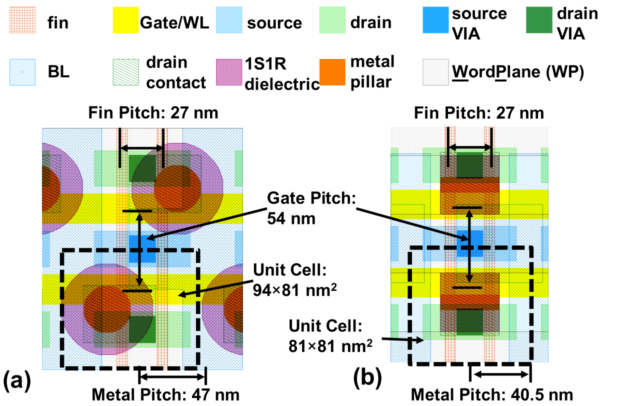

Fig. 2. Layout of pillars in (a) hexagon and (b) comb arrays using two-fin transistors.

II. From Array to Architecture

A. Array

We design both 3-D arrays (hexagon array: continuous metal plane [4] and comb array: chopped metal plane array [2], [3]) for achieving high pillar areal density (see Fig. 1). To guarantee random access to each cell, both arrays require that one transistor connects to one pillar. The pillar's areal density is determined by the larger dimensions of the MH pitch and the transistor pitch. For the hexagon array, we define the dimension of MH (D_{MH}) as the diameter of the pillar with memory dielectric. The MH pitch is assumed as $1.52D_{\mathrm{MH}}$ [15]. For the comb array, we define D_{MH} as the width of the trench. The MH pitch is $2D_{\mathrm{MH}}$. Fig. 2 shows the layout for both arrays using 7-nm FinFET design rules [17]. For the hexagon array, we use a regular hexagonal layout and the unit cell size of the pillar is 81×94 nm^2 [see Fig. 2(a)]. The gate pitch (27 nm) and fin pitch (54 nm) are given by the transistor design rules. The metal pitch is 47 nm for the regular hexagonal layout, which is larger than the pitch of 36 nm provided by the design rules. Using single-fin transistors and elliptical pillars, we can further reduce the metal pitch to the 36-nm minimum allowed by the design rule, thereby yielding the minimum unit cell size, 81×72 nm^2. The width of the minimum unit cell is determined by the metal pitch because the bitlines (BLs) need to bypass both the source and drain contacts. In the comb array, the unit cell size of the pillar is 81×81 nm^2 [see Fig. 2(b)] when two-fin transistors are used. The gate pitch and the fin pitch are the same as those in the hexagon array. The metal pitch is 40.5 nm, which can be similarly reduced to 36 nm by using single-fin transistors, yielding the minimum unit cell size, 81×72 nm^2. This also requires changing the size of the pillar from a square (40.5×40.5 nm^2) to a rectangle (40.5×36 nm^2).

The size of the transistor determines the array density. Current lithography and simple material stacks of RSMs allow

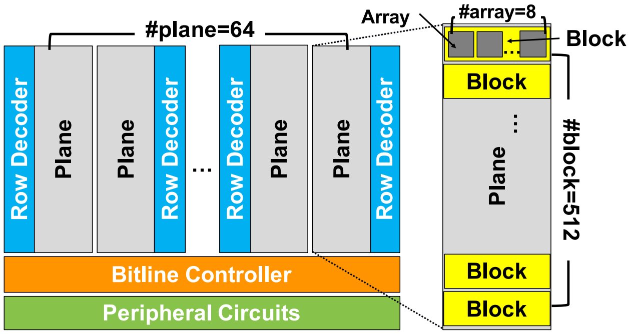

Fig. 3. Floor plan of a 3-D 64-layer VRSM chip of 1 T, 4.6 Gb/mm^2.

the scaling of the MH dimension, D_{MH}, down to sub-100 nm. Hence, the transistor size sets the unit cell size, limiting the maximum pillar density when D_{MH} is above ~100 nm. When D_{MH} falls below sub-100 nm, the etch process of pillars also restricts the scaling of the unit cell when integrating multiple WP layers (e.g. 64 layers). To minimize the size of the transistor, two column-adjacent transistors share the same source (see Fig. 2). When the transistor is scaled down to sub-10 nm node, the chip density can be larger than 4.3 Gb/mm^2—the density of 3-D NAND currently in production [13]—enabling high-density storage.

A pillar driver with small footprint and enough drive current (equivalent to 7-nm FinFET) is essential for high-density storage. Small transistor footprint increases the pillar areal density, and enough driver current is needed for successful write/read operations. The drive current of 7-nm FinFET, $\mathrm{I_{DRIVE}}$, is ~27 μA/fin [17], which is enough for LRS > 20 kΩ (LRS $> \mathrm{V_{DD}}/n\mathrm{I_{DRIVE}}$, assuming $\mathrm{V_{DD}} = 1$ V and $n = 2$, where n is the number of fins per pillar driver). Larger LRS reduces the requirement for the drive current. Here, we use the reported characteristics of high-performance logic transistors to estimate the drive current. These transistors use low $\mathrm{V_{DD}}$, a low threshold voltage, and have high OFF-current. Transistors as pillar drivers for memory applications need to handle higher $\mathrm{V_{DD}}$ with longer gate length and gradual drain junction to avoid breakdown; the threshold voltage also needs to be set higher to reduce OFF-state leakage current. Net effects of such transistor designs are not clear, but the drive current is of similar order of magnitude.

In the following analyses, we use $D_{\mathrm{MH}} = 62$ nm for both arrays, because the MH dimension D_{MH} needs to be large enough to allow a reasonably large number (e.g., 64) of stackable WP layers and the deposition of the memory and pillar materials into MHs. We assume that the radius of the core pillar is 13 nm, the thickness of the electrode is 3 nm, the thickness of memory dielectric is 3 nm, and the thickness of the selector material is 12 nm [16]. In order to simplify the analysis, we use the regular hexagonal layout and two-fin comb layout for hexagon and comb arrays, respectively. The hexagon array achieves the same bit density as the comb array when unit cells of both arrays scale by the same factor and the hexagon and comb array have the same MH dimension D_{MH}.

B. Architecture

We organize 3-D VRSM arrays into a terabit ultrahigh-density memory architecture. An example of the 64-layer architecture is given using hexagon arrays (see Fig. 3). The architecture using hexagon array (hexagon architecture) achieves a 221.57-mm^2, 1-Tb, and 4.6-Gb/mm^2 chip, which has a higher bit density than the current 3-D 64-layer NAND in production (768 Gb, 4.3 Gb/mm^2, and 3 bit/cell). Each array has 64 WP layers and 4M memory cells. Eight arrays are horizontally connected into a block using WPCs (see Fig. 4). In each memory plane, 512 blocks are vertically aligned and each block is isolated from the others. There are 64 memory planes, each paired with a row decoder, which can arbitrarily select one wordline (WL) and one WP. WLs and WPCs run the entire horizontal length of each plane. The BL controller and additional peripheral circuits are at the bottom of the chip. The BL controller and the peripheral circuits can sense one or multiple BLs. The row decoders and the BL controller guarantee random bit access. The architecture using comb arrays (comb architecture) shares a similar floor plan. The only difference is that the comb architecture [see Fig. 1(b)] requires each row decoder per memory plane to drive both sets of WPs in each array.

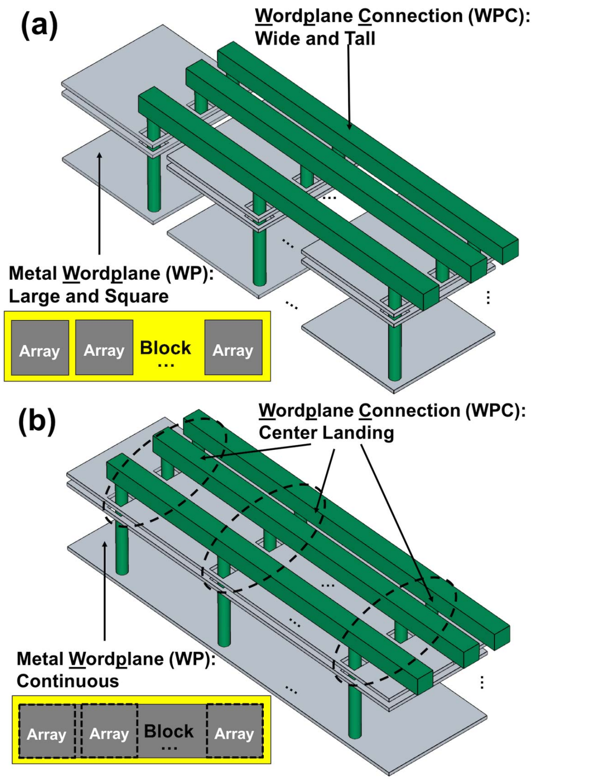

Fig. 4. Schematic of block organization using hexagon arrays. (a) Simplified block organization for simulation. (b) Continuous block organization in practice. A low-resistance WPC connects with all arrays of a layer.

With compact staircases (see Fig. 5) [18] and WPCs, we reduce the number of the row decoders, achieving high array efficiency and high density. Fig. 4 shows the block organization using hexagon arrays. WPCs connect the arrays to a block. For comb arrays, each array in the block requires two staircases [see Fig. 1(b)]. Each staircase is to contact one set of WPs in each array. The compact staircase can be achieved using the "MiLC" process, reducing the etch steps and cost [18]. The area of the compact staircase is quite small and takes less than 2% of the total chip area to allocate the

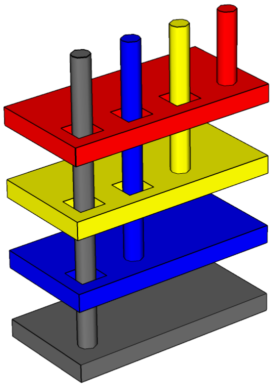

Fig. 5. Schematic of compact staircase.

TABLE I
RESISTANCES (R) AND RESISTIVITIES (ρ) USED IN THE SIMULATION OF SECTION III

	Hexagon		Comb	
	R (Ω)	ρ (Ω·μm)	R (Ω)	ρ (Ω·μm)
WP-1	54.4	0.2	80	0.2
WP-2	27.2	0.2	n/a	n/a
WP#	40	0.2	40	0.2
P	0.57	0.042	0.2	0.042
WL*	7.5	0.064	9.9	0.064
BL*	1.3	0.035	0.46	0.029

Definitions of WP-1, WP-2, WP, P, and BL can be found in Part I of the paper [16]. R_{WP-1} for comb arrays is the resistance of the arm of the WP per unit cell along the direction of the arm. R_{WPC} and ρ_{WPC} are the resistance and resistivity of WPC per unit length of array in each block. The resistivity of WPC is 0.025 Ω·μm [19]. The unit resistance of WPC, R_{WPC}, is defined as the resistance of WPC per array. It depends on the numbers of WLs and WPs. More WLs and WPs reduce the width of WPC.

#R_{WP} and ρ_{WP} are the sheet resistance and resistivity of WP.

*R_{WL} and R_{BL} are calculated under the assumption that the aspect ratio of the WL and BL is 2.

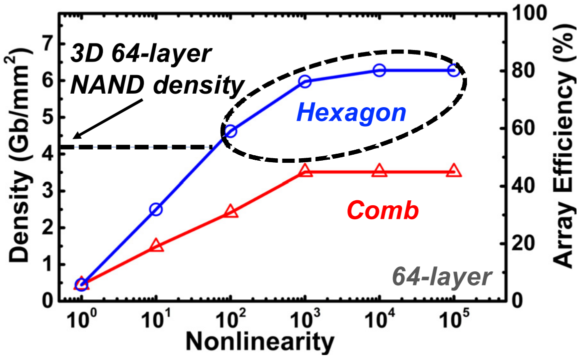

Fig. 6. Device requirements for 64-layer hexagon and comb arrays. Hexagon array with LRS = 1 MΩ and NL > 10^2 (circled) yields is required for terabit class 3-D 64-layer VRSM with >4.3 Gb/mm^2 density, denser than 3-D 64-layer NAND.

contacts of WPCs. WPCs are wide and thick wires with low resistivity (0.025 Ω•μm [19]), connecting arrays in each block in parallel and enabling more arrays in each block. In order to achieve high array efficiency, we also need to reduce the areas of staircases, row decoders, BL controller, and additional peripheral circuits. To make a fair comparison, we assume the same areas of each row decoder, BL controller, and additional peripheral circuits for 3-D NAND [11] and for 3-D VRSM. Therefore, we use WPCs wires, increase the number of bits per functional block, and reduce the number of the decoders. This design reduces the total area of all the peripherals and achieves high array efficiency (array area divided by chip area) and high chip density.

In practice, the staircase contacts in hexagon arrays can be located in the center of the arrays; it is not necessary to physically isolate WPs of the arrays on the same layer in each block—i.e., WPs can be continuous [see Fig. 4(b)]. These two modifications further reduce parasitic resistances to the worst case cell, improve the write/read margins, and provide more tolerance to the process variations. It is complex and inefficient to accurately simulate one block with the modifications, so we simplify the analysis. Here, we use the top block organization [see Fig. 4(a)], simulate the worst case behaviors of one array, extract the equivalent resistances of one array, and construct an equivalent circuit network to estimate the worst case cell behaviors in one block. The simulated results are more pessimistic and provide a reasonable device requirement guideline.

III. DEVICE REQUIREMENTS

In Part I of the article, we described a high-accuracy reduced resistor network [16]. The network is also applicable to the comb array with a relative error less than 1.0%. Here, we investigate the maximum chip density of both arrays in the chip architecture as a function of the LRS resistance value and the NL of the selector. The definitions of LRS and NL are given in Part I [16]. The resistance window [high-resistance state (HRS)/LRS] is assumed to be 10. HRS is the resistance of high-resistance state. In the simulation, we use the resistances and resistivities listed in Table I, assuming no gate leakage current in the pillar driver. Worst case single-bit write and parallel read (read all BLs simultaneously) are performed. Fig. 1 shows the dimensions of both arrays used in the simulation. The bias conditions and the criteria for write and read are summarized in Part I of this article [16]. The write/read schemes can be optimized according to specific device characteristics.

The hexagon array with large LRS and adequate NL is preferable for high-density storage (see Fig. 6). The 64-layer hexagon architecture with LRS = 1 MΩ and NL = 10^2 yields the density of 4.6 Gb/mm^2. With the same LRS and larger NL, the maximum density of the 64-layer comb architecture can only reach 2.4 Gb/mm^2. The higher density of the hexagon architecture is achieved due to large LRS, adequate NL, and its conductive hexagonal pattern. Large LRS reduces relative voltage drop on the parasitic resistances (WP, pillar, and BL). Adequate NL reduces the leakage currents through the half-selected and unselected cells in the arrays. When NL > 10^3, the achievable chip densities of both arrays saturate. Parasitic resistances draw a portion of the supply voltage and limit maximum chip density even for large NL. The continuous WP pattern in hexagon arrays reduces the parasitic resistance

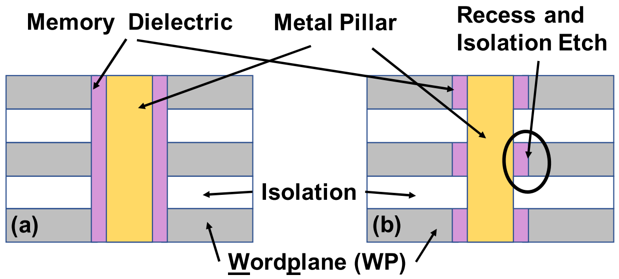

Fig. 7. Cross section of pillar. (a) Without and (b) with recess and isolation etch. Memory dielectric is not vertically connected due to recess and isolation etch process [23].

of WP, while chopped WP in comb arrays significantly increases the resistance of the WP. The maximum density of the hexagon-array architecture is 6.3 Gb/mm^2, 1.5× denser than 3-D NAND. The LRS and NL requirements stay the same when scaling V_{DD}, as long as the pillar driver can provide enough drive current. The current through the pillar driver also scales with V_{DD}.

The requirements are applicable to various types of RSMs (RRAM [2]–[4], CBRAM [5], and PCM [6]) and selectors (threshold switching [20], nonlinear [21], and diode-type [22]). To avoid the leakage current through the memory dielectric on the sidewalls of MHs, the recess and isolation etch process [23] can be used (see Fig. 7). Memory cells are then confined between WP isolation layers.

IV. Discussion

Here, we further analyze the maximum acceptable parasitic resistances of WP and pillar for various LRS and NL (see Figs. 8 and 9) with the goal of achieving bit density larger than 3-D 64-layer NAND (see Fig. 3). We assume that the array size of both hexagon and comb arrays is 64 × 256 × 256. R_{WP} and R_P are used as measures of the acceptable parasitic resistances, where R_{WP} is the sheet resistance of the WP and R_P is the resistance of the pillar per layer. Here, the unit resistances of WPC, WL, and BL are assumed with the same reasonable values ($R_{WPC} = 2.9\ \Omega$, $R_{WL} = 7.5\ \Omega$, and $R_{BL} = 0.4\ \Omega$) at the designed MH dimension for both array types (see Fig. 1). R_{WPC} is defined as the resistance of WPC per array in each block. R_{WL} is defined as the resistance of WL per unit cell along the WL direction. R_{BL} is defined as the resistance of BL per unit cell along the BL direction. With the same aspect ratios of WPC, WL, and BL, R_{WPC} and R_{WL} of hexagon array are smaller but R_{BL} of hexagon array is larger than that of comb array when the sizes of the arrays and the MH dimensions are the same. Larger R_{WPC} and R_{BL} lower write/read margins. The write/read margins are defined in Part I of this article [16]. R_{WL} barely influences the write/read margins. Hexagon array benefits from smaller R_{WPC} and suffers from larger R_{BL}; on the contrary, comb array benefits from smaller R_{BL} and suffers from larger R_{WPC}. Net effects of R_{WPC} and R_{BL} on write/read margins depend on the conductances and aspect ratios. To simplify the analysis in Section IV, we adjust the aspect ratios of WPC, WL, and BL to achieve the same unit resistance values for both array types.

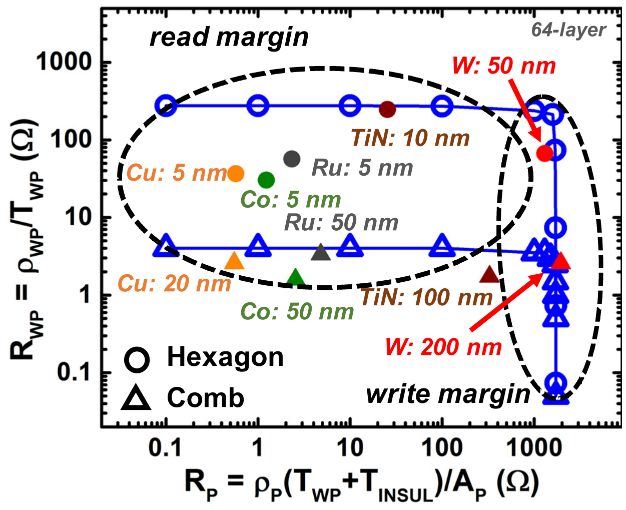

Fig. 8. Maximum resistance tolerance (blue lines) of WP and pillar for hexagon and comb arrays at LRS = 1 MΩ and NL = 10^4. Color dots with the corresponding symbols (round: hexagon and triangle: comb) indicate the resistances of WP and pillar using various BEOL materials (Cu, Co, Ru, TiN, and W).

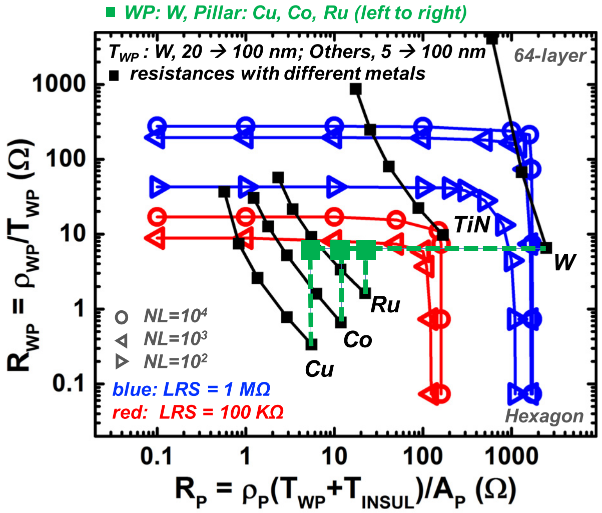

Fig. 9. Comparison between maximum resistance tolerance (red and blue lines) of WP and pillar for hexagon arrays at various LRS and NL and the resistances of WP and pillars using BEOL materials. The resistances of Cu-, Co-, Ru-, TiN-, and W-based WP and pillar (black lines) are plotted with various thicknesses of WP (left to right: 5, 10, 20, 50, and 100 nm and W-based dots: 20, 50, and 100 nm). Green dots provide an example of using different materials for WP and pillar.

To achieve ultrahigh-density 3-D VRSM, we can utilize the WP and pillar with resistances smaller than the resistances on the maximum resistance tolerance (blue lines in Fig. 8). The hexagon array (blue line with round symbols) allows for larger parasitic resistance of WP, while both arrays have the same maximum tolerance of the parasitic resistance of the pillar. Compared with the comb array (blue line with triangle symbols), the hexagon array sufficiently reduces the total parasitic resistance of WP, yielding a better tolerance of R_{WP}. The same maximum R_P is reached for both hexagon and comb arrays because the patterns of both arrays are ignored when R_{WP} approaches 0 Ω. At the region of large R_{WP} and small R_P, the boundary of the maximum resistances is determined by the read margin, where R_{WP} is the dominant factor. Here, we use parallel read, reading all BLs simultaneously. During parallel read, the total current through the selected WP is the sum of the currents of all the selected cells through the selected pillars. The total current through the selected WP is significant, much larger than the current during write. Larger R_{WP} increases the relative voltage drop on the selected WP and notably reduces the read margin. Segmented read (not reading all BLs simultaneously, 1/2 BLs, 1/4 BLs, and so on) can reduce the total current through the selected WP and endure more R_{WP}, trading off bandwidth and latency. The unselected BLs are at 0.5 $\mathrm{V_{DD}}$. At the region of small R_{WP} and large R_P, R_P is the dominant factor, and the design space boundary is controlled by the write margin. The current through the selected pillar is larger during write than read. The pillar takes up a significant relative voltage drop, and the voltage drop across R_P reduces the write margin distinctly. Assuming that the thickness of WP insulation T_{INSUL} is 6 nm [9], the typical resistances of WP and pillar using the same back-end-of-line (BEOL) materials (Cu [19], Co [24], Ru [25], TiN [26], and W [27]) with different thicknesses of WP are plotted on Fig. 8 (color dots). The hexagon array allows the scaling of WP thickness to less than 5 nm using Cu, Co, and Ru, and can also use thicker TiN (10 nm) and W (50 nm). Thicker WPs increases the difficulty in etching the 3-D structure. The comb array requires WP thickness of more than 20 nm (Cu: 20 nm, Co: 50 nm, Ru: 50 nm, and TiN: 100 nm). Both W-based WP and pillar may not be used for the comb architecture because it is out of the design space boundary, given the reported resistivity [27]. However, we can use W-based WP and other material-based pillar at 200-nm WP thickness. Considering that vertical scaling is critical for further increasing the density [9], hexagon architecture is favorable for high-density storage.

We compare the maximum resistance tolerance of different NL and LRS (blue and red lines) with the resistances of WP and pillars using the BEOL metals (black lines) in Fig. 9. Higher NL and larger LRS allow for larger parasitic resistances of WP and pillar. Higher NL reduces the leakage currents through the unselected and half-selected cells. Larger LRS increases the relative voltage drop on the selected cells. Both increase the write and read margins. The typical resistances of WP and pillar using the same BEOL materials (Cu, TiN, Ru, Co, and W) with various thicknesses (black lines) are also included here. Higher conductivities of WP and pillar with adequate thicker WP can accept lower NL and smaller LRS for the devices. Similarly, we can also estimate whether the resistances of WP and pillar using different materials fall in a design space boundary. For example, we can draw a horizontal line from the W data point at 100-nm WP thickness and three vertical lines from Cu, Co, and Ru data points at the same WP thickness (green lines). We can then infer the resistances (green dots where green lines meet) of W-based WP and Cu-, Co-, or Ru-based pillar at 100-nm WP thickness. W-based WP and Cu, Co-, or Ru-based pillar at 100-nm WP thickness are sufficient for ultrahigh-density 3-D VRSM with NL = 10^3 and LRS = 100 KΩ.

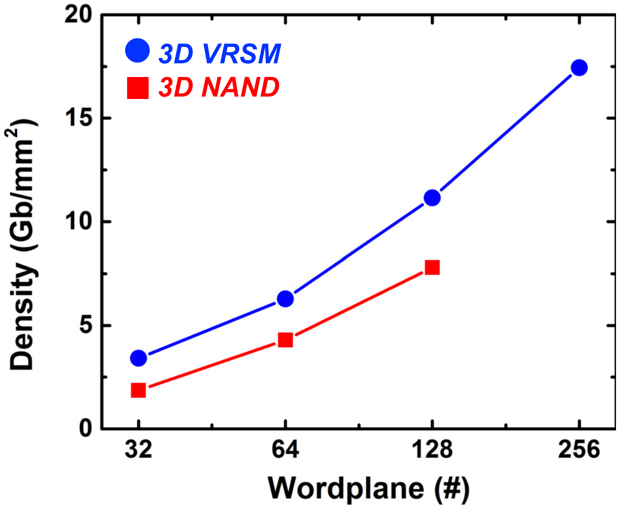

Fig. 10. Projection of 3-D VRSM with different WP layers. With the same number of WPs, 3-D VRSM has larger chip density than the reported 3-D NANDs [13], [14], [28].

A. 3-D VRSM: Hexagon Versus Comb

We compare the hexagon and comb architectures, assuming the same array size, MH dimension (D_{MH}), thicknesses (T_{OX}, T_{WP}, and T_M), and devices (NL, LRS, and HRS). In terms of fabrication, the hexagon architecture has a simpler WP pattern but requires transistors with a smaller footprint. Because of the same D_{MH}, the area of the unit cell of the pillar in the hexagon array is half of that in the comb array. The same D_{MH} (62 nm) is chosen to allow etching more than 64-layer WPs and provide enough space for metal pillar deposition. The bit density of both types is still the same. Each pillar in the comb array also has two sidewalls of devices, doubling the number of devices per unit cell. In terms of chip density, comb type suffers from requiring two staircases per array and two row decoders per memory plane, requiring more conductive WP and pillar to achieve a similar chip density than hexagon type. In terms of latency, the latency in hexagon arrays is smaller than that in comb arrays. Hexagon arrays benefit from the smaller WP resistance. The total resistance of each WP in comb arrays is 22× more than that in hexagon arrays, and the total capacitance of WP and devices in hexagon arrays is ~1.2× more than that of comb arrays. In terms of energy consumption, both types would have a similar energy consumption, since parasitic circuits consume most of the energy. Based on all these aspects, we suggest using hexagon arrays for high-density storage.

B. 3-D VRSM Versus 3-D NAND

Compared to 3-D NAND, the smaller unit cell size and compact staircase in 3-D VRSM guarantee larger chip density. We provide a projection of 3-D VRSM with different WPs in Fig. 10. With the same number of WP, 3-D VRSM can achieve larger chip density than the most advanced 3-D NANDs [13], [14], [28]. 3-D VRSM may have four material stacks in the MH, but 3-D NAND has nine stacks [29]. Less material stacks

TABLE II
SPECS OF CHIP DESIGN

	3D VRSM	3D NAND
Capacity (Gb)	512	1024
Storage Cells (#)	171	1024
Bit/Cell (b/#)	3	1
Array Area (mm^2)	106.2*	130.9
Decoder Area (mm^2)	4.2	8.3
Controller Area (mm^2)	11.5	11.5
Peripheral Area (mm^2)	10.2	10.2
Spacing (mm^2)	included in *	0.1
Staircase Area (mm^2)	included in *	2
Total Area (mm^2)	132	163
Chip Density (Gb/mm^2)	6.3	3.9

TABLE III
SPECS OF TOTAL ARRAY REGION

#WP=64	3D VRSM	3D NAND
Width/Column (nm)	94	~ 190
#Column (BL Controller Input)	2^{17}	$2^{17}\times2$
#Column	$2^{17}/1$	$2^{17}\times2/4$
Height/Row (nm)	81	~ 190
#Decoder Output	64×2^{17}	$64\times2^{15}/3$
#Row	2^{17}	$2^{15}/3\times4$
Total Width (nm)	1.2×10^7	1.2×10^7
Total Height (nm)	1.1×10^7	8×10^6

reduce the MH dimension. However, it might be difficult to etch 3-D VRSM because of its dielectrics and metals. Meanwhile, different from 3-D NAND [29], WPs of 3-D VRSM are much wider (more pillars along the BL direction per WP, e.g., 256, instead of 4). It allows us to utilize the compact staircase on the array. We assume the same areas of each row decoder, the BL controller, and additional peripheral circuits for 3-D NAND and 3-D VRSM. Though there are more row decoders in 3-D VRSM, 3-D VRSM is still denser than 3-D NAND (see Table II) overall. In this analysis, we only assume 1 bit/cell in 3-D VRSM, while 3-D NAND is assumed to be 3 bit/cell. With multilevel cell programming, 3-D VRSM can achieve even denser storage.

To better understand the array regions of 3-D VRSM and 3-D NAND, we further estimate the total widths and heights of the total array regions for both 3-D VRSM and 3-D NAND (see Table III). The 3-D NAND and 3-D VRSM have the same total width, and the total height of 3-D VRSM is slightly larger than that of 3-D NAND, which is due to the double capacity of 3-D VRSM. According to [11], we estimate the numbers of 3-D NAND in Tables II and III.

The 3-D VRSM is BEOL-friendly. If low-temperature-fabricated FETs [30]–[32] become equivalent to 7-nm FinFET in terms of drive current and transistor size, we can put the pillar transistors above the memory array and fold the row decoders, BL controller, and additional peripheral circuits underneath memory planes in 3-D VRSM. Therefore, we can stack multiple layers of memory planes for more density.

V. CONCLUSION

Using the simulation platform developed in Part I of this article [16], we present the design guidelines of device, array, and architecture to achieve ultradense 3-D VRSM. On the device level, small transistors with enough drive current (e.g., 7-nm FinFET) enable high-density storage. Large LRS increases the relative voltage on the selected cells. Adequate NL reduces the leakage currents through unselected and half-selected cells. On the array level, compared to the comb array, the hexagon array has smaller parasitic resistance of WP and is preferred in terms of write/read, latency, and energy consumption. On the architecture level, WPC and compact staircase reduce the nonstorage area and increase the array efficiency. Hexagon-array architecture with LRS $= 1$ MΩ, NL $\sim 10^3$, and only single bit per cell can achieve 1-Tb, 6.3-Gb/mm^2 3-D 64-layer VRSM. Compared to 3-D 64-layer NAND, the capacity and density increase by 30% and 46%, respectively. The 3-D VRSM with more layers of WPs can further increase the chip density (128-layer: 11.1 Gb/mm^2 and 256-layer: 17.4 Gb/mm^2). With multilevel cell programming and optimizations (low-temperature-fabricated or single-fin pillar drivers, center landing of compact staircases, continuous WPs, and write/read schemes), 3-D VRSM has good potential for future ultradense storage.

ACKNOWLEDGMENT

The authors would like to thank T. Maeda from Kioxia Corporation, A. V.-Y. Thean from the National University of Singapore, H.-Y. Chen from GigaDevice Semiconductor, L. Liyanage from Western Digital, G. Hills from the Massachusetts Institute of Technology, C.-S. Lee from Apple Inc., K.-M. Persson from Lund University, and C. M. Neumann from Stanford University for the discussions.

REFERENCES

[1] H.-S. P. Wong *et al. Stanford Memory Trends*. Accessed: Aug. 11, 2019. [Online]. Available: https://nano.stanford.edu/stanford-memory-trends

[2] I. G. Baek *et al.*, "Realization of vertical resistive memory (VRRAM) using cost effective 3D process," in *IEDM Tech. Dig.*, Washington, DC, USA, Dec. 2011, pp. 31.8.1–31.8.4, doi: 10.1109/IEDM.2011.6131654.

[3] W. C. Chien *et al.*, "Multi-layer sidewall WO$_X$ resistive memory suitable for 3D ReRAM," in *Proc. Symp. VLSI Technol. (VLSIT)*, Honolulu, HI, USA, Jun. 2012, pp. 153–154, doi: 10.1109/VLSIT.2012.6242507.

[4] H.-Y. Chen, S. Yu, B. Gao, P. Huang, J. Kang, and H.-S. P. Wong, "HfO$_X$ based vertical resistive random access memory for cost-effective 3D cross-point architecture without cell selector," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2012, pp. 20.7.1–20.7.4, doi: 10.1109/IEDM.2012.6479083.

[5] Q. Luo *et al.*, "Demonstration of 3D vertical RRAM with ultra low-leakage, high-selectivity and self-compliance memory cells," in *IEDM Tech. Dig.*, Washington, DC, USA, Dec. 2015, pp. 10.2.1–10.2.4, doi: 10.1109/IEDM.2015.7409667.

[6] K. Kurotsuchi *et al.*, "2.8-GB/s-write and 670-MB/s-erase operations of a 3D vertical chain-cell-type phase-change-memory array," in *Proc. Symp. VLSI Technol.*, Kyoto, Japan, Jun. 2015, pp. T92–T93, doi: 10.1109/VLSIT.2015.7223705.

[7] Y. Deng *et al.*, "Design and optimization methodology for 3D RRAM arrays," *IEDM Tech. Dig.*, Washington, DC, USA, Dec. 2013, pp. 25.7.1–25.7.4, doi: 10.1109/IEDM.2013.6724693.

[8] L. Zhang, S. Cosemans, D. J. Wouters, B. Govoreanu, G. Groeseneken, and M. Jurczak, "Analysis of vertical cross-point resistive memory (VRRAM) for 3D RRAM design," in *Proc. 5th IEEE Int. Memory Workshop*, Monterey, CA, USA, May 2013, pp. 155–158, doi: 10.1109/IMW.2013.6582122.

[9] S. Yu *et al.*, “3D vertical RRAM—Scaling limit analysis and demonstration of 3D array operation,” in *Proc. Symp. VLSI Technol.*, Kyoto, Japan, Jun. 2013, pp. T158–T159.

[10] L. Zhao *et al.*, “Ultrathin (~2nm) HfO_X as the fundamental resistive switching element: Thickness scaling limit, stack engineering and 3D integration,” in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2014, pp. 6.6.1–6.6.4, doi: 10.1109/IEDM.2014.7046998.

[11] R. Yamashita *et al.*, “A 512Gb 3b/cell flash memory on 64-word-line-layer BiCS technology,” in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2017, pp. 196–197, doi: 10.1109/ISSCC.2017.7870328.

[12] C. Kim *et al.*, “A 512Gb 3b/cell 64-stacked WL 3D V-NAND flash memory,” in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2017, pp. 202–203, doi: 10.1109/ISSCC.2017.7870331.

[13] T. Tanaka *et al.*, “A 768Gb 3b/cell 3D-floating-gate NAND flash memory,” in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Jan./Feb. 2016, pp. 142–144, doi: 10.1109/ISSCC.2016.7417947.

[14] C. Siau *et al.*, “A 512Gb 3-bit/cell 3D flash memory on 128-wordline-layer with 132MB/s write performance featuring circuit-under-array technology,” in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2019, pp. 218–220, doi: 10.1109/ISSCC.2019.8662445.

[15] Z. Jiang *et al.*, “Selector requirements for tera-bit ultra-high-density 3D vertical RRAM,” in *Proc. Symp. VLSI Technol. (VLSIT)*, Honolulu, HI, USA, Jun. 2018, pp. 107–108, doi: 10.1109/VLSIT.2018.8510689.

[16] S. Qin *et al.*, “Next-generation ultrahigh-density 3-D vertical resistive switching memory (VRSM)—Part I: Accurate and computationally efficient modeling,” *IEEE Trans. Electron Devices*, vol. 66, no. 12, pp. 5139–5146, Dec. 2019, doi: 10.1109/TED.2019.2950606.

[17] *ASAP ASU 7 nm PDK*. Accessed: Aug. 24, 2017. [Online]. Available: http://asap.asu.edu/asap/

[18] S.-H. Chen *et al.*, “A highly scalable 8-layer Vertical Gate 3D NAND with split-page bit line layout and efficient binary-sum MiLC (minimal incremental layer cost) staircase contacts,” in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2012, pp. 2.3.1–2.3.4, doi: 10.1109/IEDM.2012.6478963.

[19] C.-S. Lee, B. Cline, S. Sinha, G. Yeric, and H.-S. P. Wong, “32-bit Processor core at 5-nm technology: Analysis of transistor and interconnect impact on VLSI system performance,” in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2016, pp. 28.3.1–28.3.4, doi: 10.1109/IEDM.2016.7838498.

[20] H. Yang *et al.*, “Novel selector for high density non-volatile memory with ultra-low holding voltage and 10^7 on/off ratio,” in *Proc. Symp. VLSI Technol.*, Kyoto, Japan, Jun. 2015, pp. T130–T131, doi: 10.1109/VLSIT.2015.7223716.

[21] W. Lee *et al.*, “Varistor-type bidirectional switch (J_{MAX}>10^7 A/cm^2, selectivity~10^4) for 3D bipolar resistive memory arrays,” in *Proc. Symp. VLSI Technol. (VLSIT)*, Honolulu, HI, USA, Jun. 2012, pp. 37–38, doi: 10.1109/VLSIT.2012.6242449.

[22] B. Govoreanu, L. Zhang, and M. Jurczak, “Selectors for high density crosspoint memory arrays: Design considerations, device implementations and some challenges ahead,” in *Proc. Int. Conf. IC Design Technol. (ICICDT)*, Leuven, Belgium, Jun. 2015, pp. 1–4, doi: 10.1109/ICICDT.2015.7165872.

[23] K. Parat and C. Dennison, “A floating gate based 3D NAND technology with CMOS under array,” in *IEDM Tech. Dig.*, Washington, DC, USA, Dec. 2015, pp. 3.3.1–3.3.4, doi: 10.1109/IEDM.2015.7409618.

[24] N. Bekiaris *et al.*, “Cobalt fill for advanced interconnects,” in *Proc. IEEE Int. Interconnect Technol. Conf. (IITC)*, Hsinchu, Taiwan, May 2017, pp. 1–3, doi: 10.1109/IITC-AMC.2017.7968981.

[25] X. Zhang *et al.*, “Ruthenium interconnect resistivity and reliability at 48 nm pitch,” *Proc. IEEE Int. Interconnect Technol. Conf./Adv. Metallization Conf. (IITC/AMC)*, San Jose, CA, USA, May 2016, pp. 31–33, doi: 10.1109/IITC-AMC.2016.7507650.

[26] J. Bonitz, S. E. Schulz, and T. Gessner, “Ultra thin CVD TiN layers as diffusion barrier films on porous low-k materials,” *Microelectron. Eng.*, vol. 76, pp. 82–88, Oct. 2004.

[27] V. Kamineni *et al.*, “Tungsten and cobalt metallization: A material study for MOL local interconnects,” in *Proc. IEEE Int. Interconnect Technol. Conf./Adv. Metallization Conf. (IITC/AMC)*, San Jose, CA, USA, May 2016, pp. 105–107, doi: 10.1109/IITC-AMC.2016.7507698.

[28] J.-W. Im *et al.*, “A 128Gb 3b/cell V-NAND flash memory with 1Gb/s I/O rate,” in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2015, pp. 1–3, doi: 10.1109/ISSCC.2015.7062960.

[29] *Samsung 3D V-NAND Flash*. Accessed: Jun. 1, 2019. [Online]. Available: http://electroiq.com/chipworks_real_chips_blog/2015/12/02/a-look-ahead-at-iedm-2015/

[30] C. Ahn *et al.*, “1D selection device using carbon nanotube FETs for high-density cross-point memory arrays,” *IEEE Trans. Electron Devices*, vol. 62, no. 7, pp. 2197–2204, Jul. 2015, doi: 10.1109/TED.2015.2433956.

[31] S. H. Wu *et al.*, “Performance boost of crystalline In-Ga-Zn-O material and transistor with extremely low leakage for IoT normally-off CPU application,” in *Proc. Symp. VLSI Circuits*, Kyoto, Japan, Jun. 2017, pp. T166–T167, doi: 10.23919/VLSIC.2017.8008580.

[32] C.-C. Yang *et al.*, “Footprint-efficient and power-saving monolithic IoT $3D^+$ IC constructed by BEOL-compatible sub-10nm high aspect ratio (AR$>$7) single-grained Si FinFETs with record high Ion of 0.38 mA/μm and steep-swing of 65 mV/dec. and Ion/Ioff ratio of 8,” *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2016, pp. 9.1.1–9.1.4, doi: 10.1109/IEDM.2016.7838379.