Architecting 3D Vertical Resistive Memory for Next-Generation Storage Systems

Cong Xu†, Pai-Yu Chen‡, Dimin Niu†, Yang Zheng†, Shimeng Yu‡, Yuan Xie†
†Pennsylvania State University, {czx102,dun118,yxz184,yuanxie}@cse.psu.edu
‡Arizona State University, {pchen72,shimeng.yu}@asu.edu

***Abstract*—Resistive Random Access Memory (ReRAM) has several advantages over current NAND Flash technology, highlighting orders of magnitude lower access latency and higher endurance. Recently proposed 3D vertical cross-point ReRAM (3D-VRAM) architecture is an encouraging development in ReRAM's evolution as a cost-competitive solution, and thus attracts a lot of attention in both industry and academia. In this work, an array-level model to estimate the read/write energy and characterize the vertical access transistor is developed. We use the model to study a range of design trade-offs by tuning the cell-level characteristics and the read/write schemes. The design space exploration addresses several critical issues that are either unique to 3D-VRAM or have substantially different concerns from the 2D cross-point array design. It provides insights on the design optimizations of the array density and access energy, and several important conclusions have been reached. Then we propose multi-directional write driver to mitigate the writer circuitry overhead, and use remote sensing scheme to take full advantage of limited on-die sensing resources. The benefits of these optimizations are evaluated and validated in our macro-architecture model. With trace-based simulations, system-level comparisons between 3D-VRAM and a wide spectrum of memories are performed in mixed aspects of performance, cost, and energy. The results show that our optimized 3D-VRAM design are better than other contenders for storage memory in both performance and energy.**

I. Introduction

In the past decade, NAND flash based solid state drives (SSDs) have revolutionized the storage system landscape thanks to their smaller footprint, lower power, and orders of magnitude lower access latency than conventional hard disk drives (HDDs). The advent of several emerging non-volatile memory (NVM) technologies provides another opportunity to dramatically change the architecture of computer memory subsystems. Among them, spin-torque-transfer memory (STT-RAM), phase-change memory (PCM), and resistive random-access memory (ReRAM) are considered as promising candidates because all of them offer orders of magnitude lower access latency and higher endurance than flash. STT-RAM has demonstrated excellent endurance and superior switching speed [1], but it has larger cell size [1] and smaller capacity [2] than DRAM. Consequently, it is usually targeted as processor cache replacement. PCM that uses a bipolar or a diode as its access transistor can achieve similar density as DRAM chips [3], [4]. However, there is no evidence that the density of PCM could get close to that of NAND flash. ReRAM has shown better cell-level characteristics than PCM and flash [5], and a 32Gb prototype of 2-layer cross-point ReRAM has demonstrated its potential to build large-capacity memory chips [6]. The recent development in 3D vertical ReRAM (3D-VRAM) enables an ultra-high-density architecture as flash replacement [7]–[9]. There are two major reasons such 3D-VRAM can be a cost-competitive solution. First, its monolithic 3D multi-layer structure improves effective bit density dramatically, just as the conventional 3D horizontal ReRAM (3D-HRAM) does [6], [10]–[12]. Second, the cost overheads associated with additional layers are eliminated by the removal of some intermediate fabrication process [9], saving significant fabrication cost compared to a 3D-HRAM counterpart [7], [13].

This work is supported in part by SRC grants, NSF 1218867, 1213052, and by the Department of Energy under Award Number DE - SC0005026.

The scope and contributions of this paper can be classified into three categories from the array-, circuit-/architecture- to system- level design and optimizations.

Design space exploration: Most prior work on 3D-VRAM has focused on device-level optimizations for 3D vertical ReRAM cells [8], [9], [14], [15]. These devices have shown a wide range of cell-level characteristics such as resistance, nonlinearity, and switching current. A few studies [13], [16] have analyzed the 3D-VRAM array design using their circuit models, focusing on the scaling trend, the impact of geometry parameters, and comparisons between 3D-HRAM and 3D-VRAM. Despite the analysis they conducted, there is little in literature about the trade-offs of array design by exploring the cell-level characteristics and read/write schemes for 3D-VRAM. Without a detailed design space exploration, it is difficult to get insights into some design choices such as: *(1) Does the low resistance or nonlinearity of a cell play a more important role in the read/write margin? (2) Is single-bit or multi-bit access preferred in 3D-VRAM? (3) How to choose a proper read voltage to balance the sensing margin and disturbance probability? (4) What is the impact of the selection of access transistor (i.e. vertical or planar MOSFET) on the bit density?* One may argue that the answers to these questions could be tracked down from the design implications of planar cross-point ReRAM since some of the issues that arise in 3D-VRAM design appear to mimic the problems tackled previously [17], [18]. However, we find that the conclusions could change slightly (i.e. Question 1) or significantly (i.e. Question 2) in the 3D scenario from the case of planar structure. The rational behind these difference could be the existence of many more sneak paths in a 3D-VRAM array or the limited driveability of the vertical access transistor. Moreover, some issues (i.e. Question 3) are rarely mentioned in prior work and others (i.e. Question 4) are unique in 3D-VRAM.

Circuit/Architecture optimization: The write and read circuitry of ReRAM has to be carefully designed because its write drivers (WDs) and sense amplifiers (SAs) occupy a significant portion of footprint in an ReRAM chip. The area of these circuits do not scale as well as the cells, especially when the area of cells are reduced by multi-layer structure rather than technology scaling. Multi-directional driver design is proposed to mitigate these overheads by leveraging the flexibility in connecting plane electrodes and the relaxed constraints in the layout of WDs. The results show that by doing so, we can almost halve the total area of WDs and, at the same time, quadruple the array size while maintaining the design margins. In addition, remote sensing scheme is motivated to tackle the limited on-die sensing resource problem [6]. This technique is also introduced into our design. A macro-architecture model is built to quantify the benefits of these circuit/archtiecture optimizations.

System-level evaluation: After applying the array and circuit/architecture optimizations, our optimized 3D-VRAM design is compared against conventional 2D NAND flash and emerging 3D NAND flash [19], [20]. The trace-based simulations are performed by customizing an disk simulator to characterize ReRAM timing

978-1-4799-6278-5/14/$31.00 ©2014 IEEE

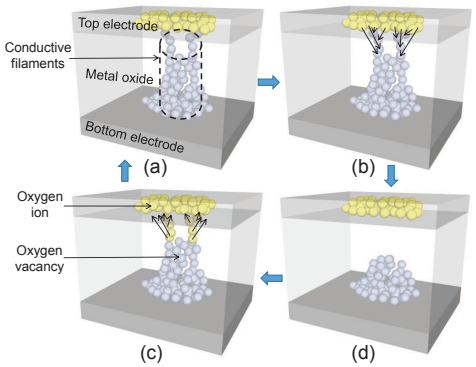

Fig. 1. Structure of an ReRAM cell and its SET/RESET operations: (a) LRS of an ReRAM cell, (b) RESET operation, (c) SET operation, (d) HRS of an ReRAM cell.

and energy models. Leveraging the cost model, more generalized comparisons can be done beyond the performance-only metrics such as input/output operations per second (IOPS) that is the focus of most system-level research on storage memory [21], [22]. In particular, the mixed performance/cost/energy metrics are interesting to the memory industry about the adoption of new memory technology. The results show that our optimized 3D-VRAM design has better IOPS/$ than other contenders for storage memory in most cases and has the best IOPS/$/J in all tested cases.

II. Preliminaries

In this section, the basics of ReRAM technology is introduced. Then we will discuss the cross-point array structure and design concerns related to its read/write operations. The 3D ReRAM architecture and design are also presented.

A. Cell Basics

As shown in Figure 1a, the basic structure of an ReRAM cell is called metal-insulator-metal (MIM): one metal oxide layer sandwiched by the top electrode (TE) and the bottom electrode (BE). Similar to PCM and STT-RAM, the state of an ReRAM cell is represented by the resistance value of the cell. The switchings between low resistance state (LRS) and and high resistance state (HRS) are caused by the formation and rupture of the conductive filaments (CFs) in the metal oxide layer [5]. A SET operation, as illustrated in Figure 1c, switches the cell from HRS to LRS. During the SET operation, a positive voltage is applied across the cell. Conductive filaments (CFs) made of oxygen vacancies are formed due to the electrical field. Upon the completion of SET process, the cell becomes LRS, as shown in Figure 1a. The RESET operation, as illustrated in Figure 1b, is a reversed process of the SET operation. During the RESET operation, the oxygen ions are forced back to the metal layer and recombine with the oxygen vacancies. After RESET, the CFs are "cut off" and the cell becomes HRS (Figure 1d).

In comparison with NAND flash, this technology has demonstrated superior cell-level characteristics with better scalability ($< 10nm$). In some aspects, its advantages over flash can be orders of magnitude including faster read/write access latency ($\leq 100ns$), higher endurance (up to 10^{12}), and lower operating voltage ($\leq 3V$) [5].

B. Array Structure

1) Planar structure: There are two basic structures of a planar ReRAM array: the MOSFET-accessed (1T1R) structure and the cross-point structure. In the 1T1R structure, each ReRAM cell has a dedicated MOSFET as its access transistor. It is easy to control each cell in such structure independently with minimum crosstalk.

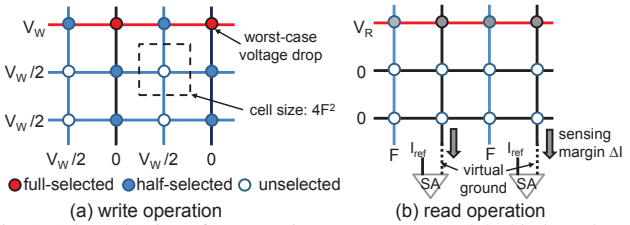

Fig. 2. Schematic view of across-point array structure and the biasing schemes of its (a) write and (b) read operation.

However, the size of the access transistor should be sized up to satisfy the current requirement of the SET/RESET operation. The total area of 1T1R ReRAM is determined by the footprint of transistors rather than ReRAM cells. On the other side, the cross-point structure is a more area-efficient approach to build an ReRAM array. As shown in Figure 2, in the cross-point structure, each ReRAM cell is located at the cross-point of a horizontal word-line (WL) and a vertical bit-line (BL). As can be seen in Figure 2a, each cell occupies a small footprint of $4F^2$ (F is the technology feature size), which is the theoretically minimum cell size for a single-layer memory structure.

When writing one or more cells in an array, the WL and BL(s) connected to the cell(s) are activated to cause enough voltage drop on selected cell(s). At the mean time, the other unselected WLs and BLs are biased properly or left floating to avoid the write disturbance. One of the most common write biasing schemes [18] is shown in Figure 2a: during the write operation, the selected WL is biased at V_W while the selected BLs are grounded. And all the unselected WLs and BLs are half biased at $V_W/2$. In the ideal case when the wire resistance is not considered, the write voltage V_W (=V_{SET} or $-V_{RESET}$) fully drop on the *full-selected cells*, and the voltage dropped on the *half-selected cells* located at the same WL or BLs with the *full-selected cells* is $V_W/2$. There should be no voltage drop across all the *unselected cells*. In practice, even with proper write schemes, the sneak current of the *half-selected cells* together with the write current of *full-selected cells* result in significant IR drop on the wire resistance, reducing the amount of voltage drop on these *full-selected cells*. In order to implement a reliable cross-point array, the worst-case voltage drop the furthest full-selected cell should be larger than the threshold write voltage given the duration of write pulse [5]. A common solution to the sneak current problem is to suppress the current of the *half-selected cells* by introducing nonlinearity into the cells. The nonlinearity K_r means the resistance of an ReRAM cell increases as the voltage across it decreases, and is defined as $K_r = R(V_W/2)/R(V_W)$, where $R(V_W)$ and $R(V_W/2)$ are the equivalent resistance values at write voltage and half write voltage respectively.

When reading one or more cells in an array, the selected WL is biased at V_R while the unselected WLs are grounded and the unselected BLs are left floating. The selected BLs are connected to current-mode sense amplifiers (SAs) and they are virtually grounded [23], as shown in Figure 2b. Then the state of the selected cells can be successfully identified by the SAs if a sufficient sensing margin is well established. The sensing margin is determined by $\Delta I = min(I_{LRS}) - max(I_{HRS})$, where $min(I_{LRS})$ is the minimum read current when the selected cell is in LRS and $max(I_{HRS})$ is the maximum read current when the selected cell is in HRS. Note that the sensing model we develop in this work is based on the current-mode sense amplifiers and ΔI is used to represent the sensing margin. Such sensing techniques are demonstrated in most prototypes of emerging NVMs [12], [24] due to its faster sensing speed. However, some prior work assumes voltage-mode sense amplifiers [17], [22] in the read operation of their cross-point models, in which the selected BLs to connected to the ground

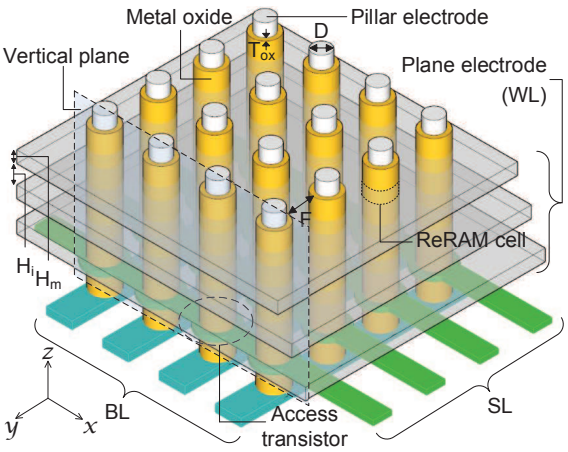

Fig. 3. Schematic view of a 3D-VRAM array with vertical access transistor

through sense resistance and ΔV is used to represent the sensing margin. They could potentially claim different conclusions.

2) Three-dimensional structure: As the cost of NAND flash continue to decline either with technology scaling or structure innovation, the key challenge for ReRAM to replace NAND Flash is to improve its integration density in terms of cost-per-bit. One simple solution is to stack the planar cross-point structures layer by layer (a.k.a 3D-HRAM), which improves the bit density [13] to $0.25L\ b/F^2$ where L is the number of layers in 3D-HRAM. However, the fabrication cost of critical lithography, etching, chemical mechanical planarization, and other process associated every stacked layer increases linearly with L. These cost adders may eventually offset the benefits of larger density enabled by the multi-layer structure [7], [13].

From the industrial perspective, the cost per bit is a major driving force to pursue the 3D integration. This motivates the 3D-VRAM architecture, which tilts the horizontal ReRAM by 90 degrees, as a much more cost-efficient solution by eliminating the aforementioned fabrication cost overheads in 3D-HRAM. The schematic view of a 3D-VRAM array is illustrated in Figure 3. The plane electrodes and isolation layers are deposited consecutively. Only after the top most layer is deposited, the critical lithography and etching steps are involved for patterning the pillar electrodes and opening the contacts for WLs. The vertical ReRAM cells are sandwiched between the perpendicular pillar electrodes and multi-layer plane electrodes. At the bottom of the pillar electrodes, there is a 2D arrays of access transistors. Their sources are connected BLs and their gates are controlled by the select-lines (SLs). With the appropriate bias schemes on WL (decoding in z-direction), BL (decoding in x-direction) and SL (decoding in y-direction), each memory cell in the 3D cross-point architecture can be individually accessed. During a read or write operation, one selected SL is biased to turn on the access transistors connected to this SL while all the other access transistors remain off by grounding the unselected SLs. This operation basically activates an x-z vertical plane, which is a de facto cross-point structure. Within the vertical plane, the same read/write biasing scheme as planar cross-point structure can be applied.

Planar access transistors (PATs) are implemented in some prior work [13]. Their results show that the bit density of many design points in 3D-VRAM are bounded by the size of PATs. To overcome the problem, the PATs are replaced with vertical access transistors (VATs) in our design. As a result, the planar footprint of an 3D-VRAM cell can be as small as $4F^2$ when it is not bounded by the etching aspect ratio, which is 33% less than the minimum planar cell size in 3D-VRAM with PATs. However, the maximum number of layers in 3D-VRAM is limited by the drivability of the VATs. Our model considers these effects.

TABLE I
SIMULATION PARAMETERS

Metric	Description	Value(s)
F	Feature size of the design	$30nm$
H_m	Height of a plane electrode	$20nm$
H_i	Height of an isolation plane	$10nm$
H_s(=$H_m + H_i$)	Height of a vertical stack	$30nm$
T_{ox}	Thickness of the switching layer	$5nm$
D	Diameter of a pillar electrode	-
P	Planar cell-to-cell pitch	-
AR	Etching aspect ratio	16
N	Number of BLs and SLs	$16 \sim 256$
L	Number of layers	16
N_b	Number of accessed bits in parallel	$1 \sim 128$
V_W	Write voltage	$3.0V$
V_R	Read voltage	$0.5V, 1.5V$
K_r	Nonlinearity of an ReRAM Cell	$5 \sim 200$
R_{on}	ON-state resistance at V_W	$25k\Omega \sim 500k\Omega$
R_{off}	OFF-state resistance at V_W	$2.5M\Omega \sim 50M\Omega$
I_{on}	Saturation current of a VAT	$100uA$

III. 3D-VRAM ARRAY DESIGN

A. Array-Level Model

A 3D sub-circuit module with ReRAM cells, interconnect resistors is implemented in HSPICE following the approach proposed in prior work [13]. The sub-circuit module is then replicated in the 3D space to simulate a full size 3D-VRAM array. The saturation current of the VATs is taken into considerations in our array model. We also model the read/write energy of a 3D-VRAM array. The geometries and other design parameters in this work are summarized in Table I.

1) Bit density: The cell-to-cell pitch is the distance from one cell to another adjacent cell in the same planar electrode,

$$P = D + 2T_{ox} + F \tag{1}$$

The etching aspect ratio (AR) is defined as,

$$AR = \frac{H_s \times L}{D + 2T_{ox}} \tag{2}$$

with the constraint that $D + 2T_{ox} \geq F$.

Then the cell area is calculated as,

$$A_{cell} = max(4F^2, (\frac{H_s \times L}{AR} + F)^2) \tag{3}$$

The bit density can be derived from Equation 3,

$$D_{bit} = \begin{cases} \frac{1}{4}L \ \ (b/F^2) & \text{if } AR \geq \frac{H_s}{F} \times L \\ \frac{1}{(\frac{H_s \times L}{F \times AR}+1)}L \ \ (b/F^2) & \text{if } AR < \frac{H_s}{F} \times L \end{cases} \tag{4}$$

As seen from Equation 4, when not bounded by the etching aspect ratio, that is, if $AR \geq L$ (because $H_s = F$ in our simulation settings), the theoretically maximum bit density $0.25L\ b/F^2$ is achieved thanks to the introduction of VATs. The AR in this work is assumed to be a modest value - 16. And most of our evaluations are based on 16-layer 3D-VRAM with VATs. It is worth mentioning that the maximum L with tolerable noise margin is limited by the saturation current of the VAT since the VAT should be able to sink the total current on a selected pillar electrode during the write operation.

2) Write and read margin: To set up appropriate values for V_W and the criterion of write margin, the switching voltage distribution within an array of cells should be considered as ReRAM are well-known for its switching parameter variability [5]. For ReRAM cells with an average switching voltage of $2V$, a possible switching range of $1.7V \sim 2.3V$ is assumed, as shown in Figure 4. In this work, the write access threshold is set to $2.5V$ to ensure a safe write operation,

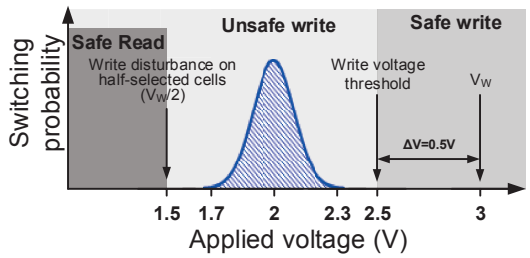

Fig. 4. Voltage regions of sate write (perfect switching probability), unsafe write (not perfect switching probability), and safe read (no disturbance).

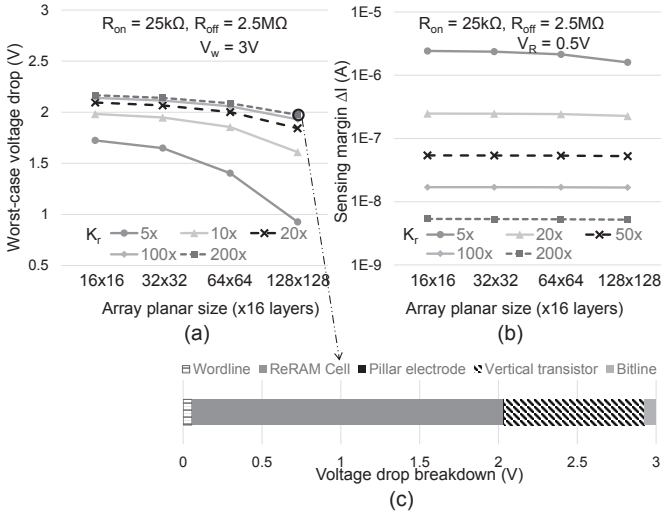

Fig. 5. A 16-layer 3D-VRAM with low resistance $R_{\text{on}} = 25k\Omega$: (a) Worst-case voltage drop on furthest cell with various nonlinearity K_r; (b) read sensing margin with various K_r; (c) voltage drop breakdown of each element in a $128 \times 128 \times 16$ array with $K_r = 200$.

and V_W is set to $3V$ to obtain a $0.5V$ toleration of voltage loss on the interconnect. Meanwhile $V_W/2 = 1.5V$ is lower than $1.7V$ to avoid the disturbance of the *half-selected cells*. For the read margin, a minimum $\Delta I = 50nA$ is used as the criterion for a reliable sensing with reasonable latency [25]. The maximum V_R is the same as $V_W/2$ to avoid the disturbance of the cells on selected WLs.

B. Design Space Exploration

The design trade-offs involving cell-level characteristics, read/write schemes will be discussed.

1) Nonlinearity: Figure 5 shows the worst-case voltage drop on the furthest selected cell and read sensing margin in a 16-layer 3D-VRAM array with various nonlinearities versus the planar size of an array. Even with a large nonlinearity of 200, the worst-case voltage drop cannot meet the criterion. This is because the write current for $R_{\text{on}} = 25k\Omega$ already exceeds the saturation current ($100\mu A$) of VAT, which causes significant voltage drop across the VAT, as illustrated in the breakdown of voltage drop in Figure 5c. On the other hand, the increase of nonlinearity dramatically decreases the read sense margin. Given R_{on}, the sensing margin ΔI decreases dramatically as the nonlinearity increases. These results suggest that current drivability of the vertical transistor put a hard constraint on the minimum R_{on}. In other words, with a small R_{on}, increasing nonlinearity alone is not able to meet the write margin criterion while it is detrimental to the read sensing margin.

2) Resistance: Figure 6a illustrates that even with a small K_r of 5 and V_R of $0.5V$, increasing R_{on} (up to $500k\Omega$) is an effective way to improve the worst-case voltage drop while maintaining the read

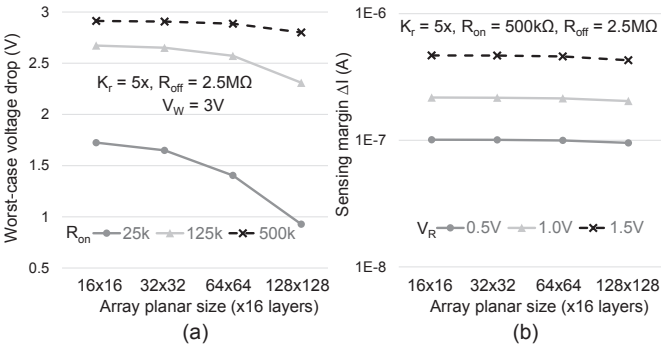

Fig. 6. A 16-layer 3D-VRAM with nonlinearity $K_r = 5$: (a) Worst-case voltage drop on furthest cell with various nonlinearity R_{on}; (b) read sensing margin with various read voltage V_R

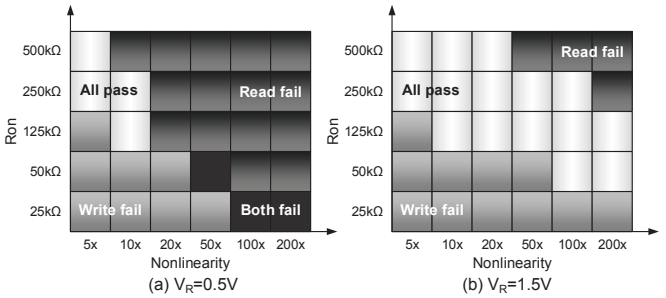

Fig. 7. Shmoo plot with different R_{on} and nonlinearity in a $128 \times 128 \times 16$ array for (a) $V_R = 0.5V$ and (b) $V_R = 1.5V$.

margin above the criterion. Along with the observations in Figure 5, it is concluded that **R_{on} plays a more important role in 3D-VRAM array than K_r does due to the limited drivability of the VAT**. Such conclusion does not apply on the planar cross-point array design [17], [18]. As a result, 3D-VRAM relaxes the design efforts from the perspective of device engineering because it is easier to increase R_{on} by lowering the SET compliance current while it is more difficult to improve K_r which typically requires metrial/structure innovations or additional selector devices.

3) Read voltage: When tuning cell-level characteristics, there could exist a fundamental conflicting nature between the write and read margin. One design knob that can be tuned is the read voltage V_R, as long as it is less or equal than $V_W/2$ to avoid the disturbance of the cells in the selected plane electrode. Figure 6b demonstrates the sensing margin with different V_R as a function of array planar sizes in a 3D-VRAM array. By increasing V_R from $0.5V$ to $1.5V$, the sensing margin are improved approximately by a factor of 4. With a larger V_R, a higher R_{on} or a larger nonlinearity can be tolerated. However, the read access energy also increase by about 10X from $0.5V$ to $1.5V$. The optimizations of access energy are presented later.

To better understand how the read voltage enables more design points in the large design space, the cell-level parameters R_{on} and K_r are swept to see the effect. Figure 7 illustrates the Shmoo plot that describes the read or/and write failure of a $128 \times 128 \times 16$ 3D-VRAM array for both $V_R = 0.5V$ and $VR = 1.5V$. Not surprisingly, a large V_R relaxes the constraints on the sensing margin, and allows more design configurations to pass the criterion.

4) Read and write energy: We develop an energy model to evaluate the read/write energy of a 3D-VRAM array. It is found that the static energy consumption due to the sneak path current, rather than the dynamic energy consumption due to the charging/discharging parasitic capacitances, dominates the access energy because a lot of sneak paths exist in an activated 3D-VRAM array. Therefore

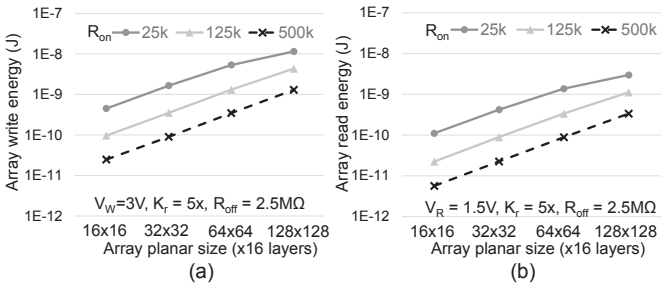

Fig. 8. Access energy of a 16-layer 3D-VRAM array with various R_{on} for (a) writing a single bit in an array; and (b) reading a single bit in an array.

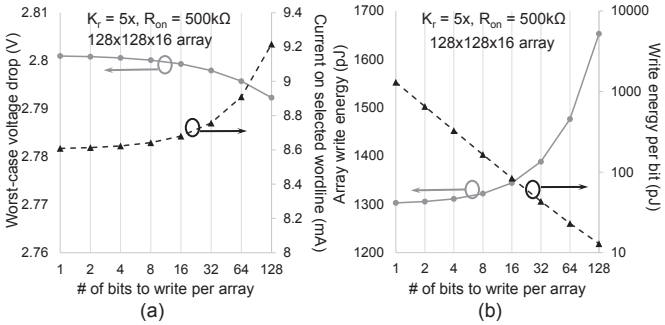

Fig. 9. Multi-bit write operation in a 128x128x16 3D-VRAM array: (a) worst-case voltage drop and write current versus number of bits to write in parallel; (b) array write energy and write energy per bit versus number of bits to write in parallel.

increasing R_{on} or nonlinearity should also reduce the access energy significantly. Figure 8 shows the read and write access energy when reading and writing a single cell in a 16-layer 3D-VRAM array. The duration of a write pulse is assumed to be $100ns$ [18], and the read sensing latency for $\Delta I \geq 100nA$ can be as small as $26ns$ [25].

5) Multi-bit access: Theoretically, the entire row of the selected WL can be read or written in parallel. In practice, only a small number of bits are accessed at the same in a planar cross-point structure. The primary reason is that the total current on the selected WL increases dramatically as the number of *full-selected cells* increases. It degrades write margin and incurs high area overheads of WDs [18]. A planar cross-point array sized by 512×512 (=256k cells) with the same cell characteristics as our 3D-VRAM cells is simulated. The current requirement of an individual WD doubles when increasing the number of accessed bits in parallel from 1 bit to 128 bits.

Does this conclusion still hold for 3D-VRAM? Figure 9a plots the write margin and write current on the selected WL as a function of the number of bits N_b that are written in parallel in a 3D-VRAM array with the same number of cells (256k) as the 2D case. The write margin degrades slightly as the N_b increases. It can also be observed that the increasing of write current from 1-bit write to 128-bit write is only 7%, suggesting that multi-bit write operation is feasible in 3D-VRAM. The rational behind the difference between 2D and 3D ReRAM is that the number of *half-selected cells* on the selected WL in the 3D-VRAM (=$16284 - N_b$) is much more than they are (=$128 - N_b$) in the 2D cross-point array, and thus the current of these *half-selected cells* dominates the total current on a selected WL.

The array write energy and write energy per bit as a function of N_b is shown in Figure 9b. We can see that the write energy of the 3D-VRAM array increases by only 28% from 1-bit write to 128-bit write, and the write energy per bit is substantially reduced as N_b increases. We also examine the read case and find similar trends and

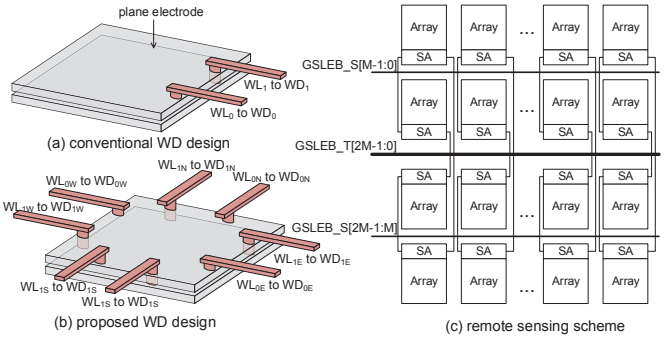

Fig. 10. Circuit optimizations for 3D-VRAM: (a) conventional write driver (WD) design: WDs are connected to one edge of the array; (b) proposed design: WDs are connected to the North, South, East, and West edges of the array; (c) remote sensing scheme: shared SAs within a block (adapted from [6] with modifications.

conclusions (not shown due to the space limit). It is concluded that **multi-bit access is much more favorable in 3D-VRAM than single-bit access with high energy efficiency and low area overheads.**

IV. Circuit and Architecture Design

In this section circuit techniques to relax peripheral overheads are introduced. Then we will explain macro-architecture design and use our developed macro model to evaluate some of the optimizations.

A. Optimize write and read circuitry

Targeted as NAND flash replacement, the ReRAM design should be highly optimized for cost-per-bit, which is primary determined by the die area of an ReRAM chip given the die capacity. Several factors have major impacts on the die area: (a) bit density determines the total area of cells, (b) array size determine the number of sets of peripheral circuits (i.e. decoders, multiplexers, write drivers etc.), (c) the style of peripheral circuity affects its area. The design space exploration in Section III tries to find optimal design points with high bit density and large array size. The techniques to be discussed in this section relaxes the peripheral overheads.

1) Multi-directional write driver: In traditional memory structures including SRAM, DRAM, flash and 1T1R NVM, the WDs in the last-stage row decoders are sized up to balance the delay of charging/discharing the corresponding WL. Alignment of these WDs is challenging because the WDs have to layout in the space of WL-defined pitch. One solution is to layout the WDs with even-numbered WLs on one side of the WLs and the WDs with odd-numbered WLs on the other side of the WLs. For cross-point NVMs, the WDs are responsible for providing sufficient current of the selected WL to both the *full-selected cells* and the *half-selected cells*. This not only worsens the alignment problem but also increases the area of these WDs, reducing the array efficiency significantly. Our 3D-VRAM design tries to solve this problem. First, the alignment problem is much alleviated in 3D-VRAM because for an $N \times N \times L$ array there are only L WDs to be aligned aside a planar size of $N \times N$, increasing the effective SL-defined pitch by N/L. For example, as shown in Figure 10a, the L WDs are connected to one edge of the array through top metal layer in a conventional design. The metal vias defined by previous fabrication process are responsible for connecting these WDs to the plane electrodes.

Utilizing the flexibility in the placement of contacts for plane electrodes, we propose multi-directional WDs for 3D-VRAM. The design is demonstrated in Figure 10b. The WD for topmost plane electrode, marked as $\text{WD}_{1\text{N}}$ in the conventional design, are distributed to the north ($\text{WD}_{1\text{N}}$), east ($\text{WD}_{1\text{E}}$), south($\text{WD}_{1\text{S}}$), and west($\text{WD}_{1\text{W}}$) of

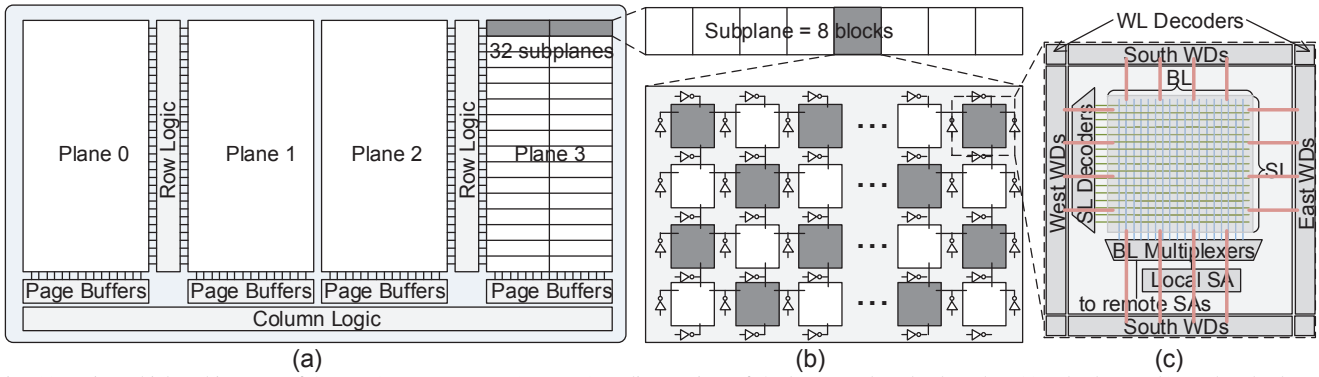

Fig. 11. Hierarchial architecture of 3D-VRAM macro: (a) An ReRAM die consists of 4 planes, and each plane has 32 sub-planes; (b) Each sub-plane is made up of 8 blocks, the SAs and WDs within a block are shared among the 3D-VRAM arrays; (c) a detailed view of a 3D-VRAM array.

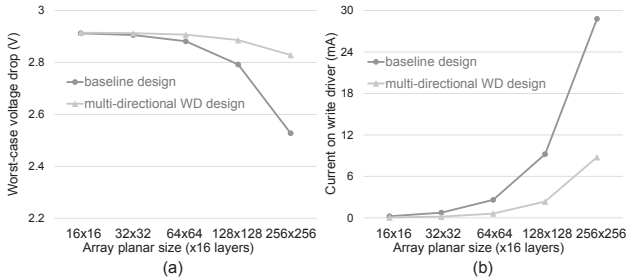

Fig. 12. Multi-directional write driver design versus conventional design in (a) Write margin of the 3D-VRAM array and (b) Current requirement of individual WD.

the array in the new design. By doing so, each single WD only need to provide one quarter of the current on that plane electrode, reducing the size of an individual WD and further relaxing the alignment constraints. Moreover, most WDs can be shared by two adjacent arrays along either the SL-direction or BL-direction, as illustrated in Figure 11b. For example, the WDs placed on the south of the array can be shared with the adjacent array to the south of it if either array is activated. Therefore, the array efficiency is improved as the total number of sets of WDs in a block is reduced.

Another side benefit of the proposed WD design is that the current path for the worst-case voltage drop on the plane electrode is almost halved, improving write margin significantly. As a result, larger array size may be allowed. Figure 12 shows the comparisons of write margin and write current between the multi-directional driver design and the conventional design. It is observed that the worst-case voltage drop of a $2N \times 2N \times 16$ array in the new design is slightly better than that of a $N \times N \times 16$ array in the baseline design. Moreover, the current requirement of each individual WD for $2N \times 2N \times 16$ arrays in the new design is almost the same as it is for $N \times N \times 16$ arrays in the baseline design. The sensing margin is well-maintained in the new design. Therefore it can be concluded that **with the proposed write driver design we can quadruple the array size and reduce the total area of WDs at the same time**.

2) Remote sensing scheme: We find that the area overhead of current-mode SAs is significant after we surveyed a broad range of state-of-the-art nonvolatile memory prototypes [2], [3], [6], [10], [12]. Our calculation shows that the layout area of a current-mode SA is in the range of $10^4 \sim 10^5 F^2$. Given the footprint of one array in our 3D-VRAM design is in the order of $10^5 F^2$, the sensing resources in an ReRAM die is very limited as the array efficiency is an important design criterion. We use the concept of remote sensing scheme introduced in a recent 3D-HRAM prototype [6]. As shown in Figure 10c, for a block with $2M$ arrays with each array having its local SA, the global select buses GSELB_S are used to control the connections between local SAs and the central buses GSELB_T. The GSELB_T of the accessed array are multiplexed into one group of buses which connect to the selected BLs in the activated array. Only one array in a block can be activated at a time. The read operation within a block is pipelined to read out the required amount of data. The parasitic delay in the cross-block buses for read operation are calculated in our macro-level model.

B. Macro-Architecture Design

The architecture of our 3D-VRAM macro is illustrated in Figure 11. Each ReRAM die is designed as a multi-plane architecture and multiple memory requests are served in parallel. Within each plane, there are 32 sub-planes and two sub-planes in the same row are activated at the same time. Each sub-plane is further divided into 8 blocks and 1 of them are activated during access. Assuming there are $2M$ arrays in a block, the M arrays (marked in dark grey in Figure 11b) are activated for writing the first half of data in the block, then the remaining M arrays (marked in white) are activated for writing the second half of data in the block. The switching between them is fast because we can simply disable the output of one direction and enable the output of the opposite direction in all the activated WDs in a block.

C. Macro-Level Model

We implement the architecture of our 3D-VRAM design in NVSim [26], which is an open source modeling framework for emerging NVMs. To evaluate the area and energy savings of our proposed design, the modules of write drivers and sensing circuitry/structure are heavily modified in NVSim.

1) Timing model: The physical access time for reading a page in 3D-VRAM can be expressed as,

$$t_{page\_read} = S_r \times t_{sense} + t_{peri} + t_{trans} \quad (5)$$

where S_r the number of serial sensing steps within a block, t_{sense} is the sensing delay including both the latency of the sense amplifiers and the RC delay of cross-block buses, t_{peri} is the delay of other peripheral circuits such as decoders and multiplexers, and t_{trans} is the data transfer latency from page buffers to I/O. Normally the internal data movement is transferred byte by byte, then the data transfer latency can be calculated by $t_{trans} = N_p / f_{trans}$ where N_p the page size and f is the data transfer frequency.

2) Cost model: We reconstruct the cost models of ReRAM from previous work [13]. One modification made is to break down the details of fabrication process of VAT include its corresponding cost overhead in the IC Knowledge LLC [27].

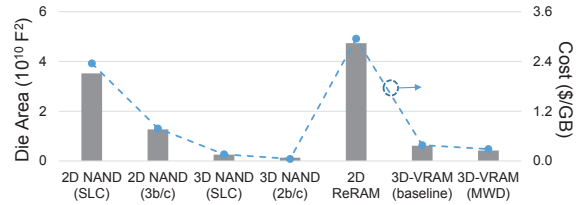

Fig. 13. Area of 64Gb (left y-axis) chip and Cost per GB(right y-axis) at $F = 30nm$ for 2D SLC NAND, 2D 3b/c MLC NAND, 16-layer SLC NAND, 16-layer 2b/c NAND, planar ReRAM, baseline 3D-VRAM, 3D-VRAM with multi-directional writer driver design.

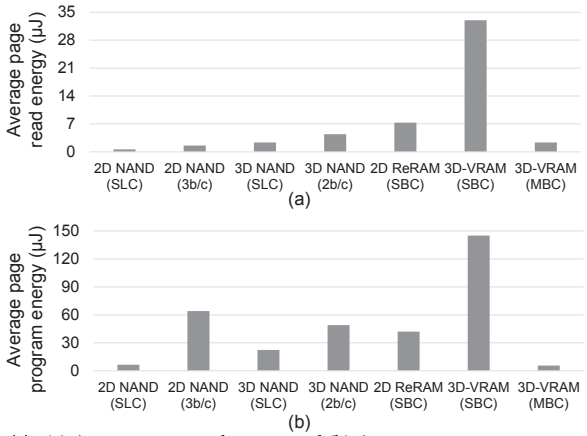

Fig. 14. (a) Average page read energy and (b) Average page program energy at $F = 30nm$ for a page size of 8KB in 2D SLC NAND, 2D 3b/c MLC NAND, 16-layer SLC NAND, 16-layer 2b/c NAND, 2D SLC NAND, 2D 3b/c MLC NAND, 16-layer SLC NAND, 16-layer 2b/c NAND, planar ReRAM with single-bit access(SBC), 16-layer 3D-VRAM with SBC, 16-layer 3D-VRAM with multi-bit access(MBC).

D. Results and Discussions

The die area of 64Gb chips with different memory organizations are compared in Figure 13. MLC ReRAM in 1T1R structure is feasible, but it is not considered in this work because the actuate control of resistance values after programming is difficult to achieve in any form or cross-point structure. We can see that the invention of 3D vertical structure in both NAND flash and ReRAM can reduce the die area substantially. ReRAM has larger die size that its NAND flash counterpart with the same bit density because the WDs and SAs are much larger than they are in NAND flash. The multi-directional WD design reduces the overall die are of 3D-VRAM from $6.1 \times 10^{10} F^2$ to $4.2 \times 10^{10} F^2$.

The cost per GB of these memories are also plotted. To make a fair comparison, the calculations are based on the same feature size $F = 30nm$ for different memory structures and organizations. And all the simulations later assume $F = 30nm$ unless specified. As seen in Figure 13, the cost comparison almost follows the trend in the die area comparison, affirming that the process of 3D vertical structures do not introduce significant cost adders.

Figure 14 compares the page access energy among different memory organizations. The 2D ReRAM with multi-bit access are not shown because its area overhead is too large due to the aforementioned reason in Section III-B5. We can see that if single-bit access is implemented in 3D-VRAM, the read energy would be much larger than other memories because it is aggregated from a large number of activated arrays which have a lot of sneak paths. With multi-bit access, the read energy could be reduced substantially. Similar conclusion applies to the write scenario. The multi-bit write operation reduces the write energy of the 3D-VRAM array from about 3X to only 12.5% that of the 3D MLC NAND.

TABLE II
TIMING PARAMETERSFOR NAND FLASH AND RERAM

Item	SLC NAND	2b/c NAND	3b/c NAND	ReRAM
Read latency (μs)	35	50	90	6.4
Program latency (μs)	350	350∼3000	350∼5000	0.5
Erase latency (ms)	1.5	5.5	10	N/A

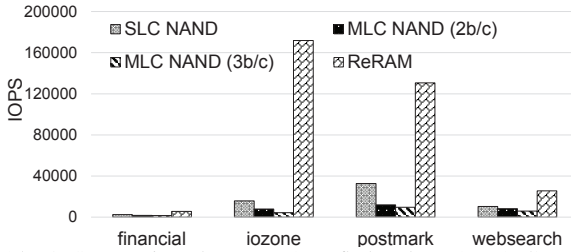

Fig. 15. IOPS comparisons between NAND flash and ReRAM (2D/3D) across different applications

V. SYSTEM-LEVEL EVALUATION

After applying the optimizations on our 3D-VRAM design, trace-based simulations are performed by customizing an disk simulator with SSD extension [21] to characterize ReRAM models. Different workloads with various I/O request frequency and patterns are simulated, including the synthetic workload in the disk simulator [21], Iozone and Postmark, as well as Financial and Websearch [28].

In the macro model of storage memories, it is observed that the t_{peri} term in Equation 5 contributes to less than 2% of the total physical page access time in most configurations. Therefore, the different peripheral delays between 2D and 3D memories in the system-level performance evaluation can be ignored. We use generalized timing parameters for SLC NAND, MLC NAND and ReRAM. Table II summarizes the page read/program latency of them and the block erase latency of NAND flash. These specifications are based on 64Gb NVM dies with 8KB page size, and the I/O date transfer rate is 166MBps.

The performance comparison between ReRAM and different NAND flash are illustrated in Figure 15. It is observed that ReRAM (2D or 3D) as storage memory can improve the system throughput greatly. The increasing of IOPS are remarkable for the workloads with high (e.g. iozone) and modest (e.g. postmark) write intensity. Performance improvement over read-intensive workloads (e.g. financial and websearch) are also significant for ReRAM.

Performance-only metric is not sufficient for evaluating the potential of a new memory technologies to be adopted in industry. A major reason that SSDs took over the storage market from HDD is that it has lower price/performance ratio than HDD. Therefore we introduce the metric of IOPS/$ to compare the emerging 3D-VRAM with the existing technology (2D NAND flash) and other contenders (e.g. 3D NAND flash). Figure 16a shows the comparison results (the IOPS/$ of every configuration for a given workload is normalized to that of the 2D SLC NAND for the given workload). Our optimal 3D-VRAM design wins over most of other memories, including its 3D SLC NAND counterpart, in all the tested workloads. For iozone and postmark, the advantages of 3D-VRAM over others are more than 45%. However, for read-intensive workloads, the IOPS/$ of 3D-VRAM can be 35% less than that of its 3D MLC NAND flash counterpart.

Another metric - IPOS/$/J - is also proposed, which combines the performance, cost, and energy aspects of a memory technology. As shown in Figure 16b, our optimal 3D-VRAM design is a clear winner over all the other memories for all tested workloads.

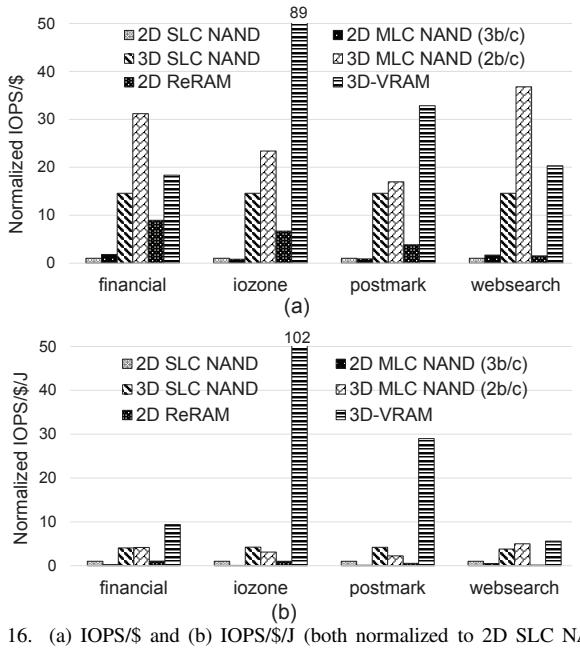

Fig. 16. (a) IOPS/$ and (b) IOPS/$/J (both normalized to 2D SLC NAND for each application)

VI. Conclusion

ReRAM is one of the most promising candidates for next-generation storage systems. Compared to NAND Flash, ReRAM has superior read/write access latency and many other advantages. 3D-VRAM has been demonstrated as a naturally low-cost architecture soluion. As changes to the existing memory technology are challenging, it is critical to study every characteristics of the new technology that could affect the design choices. We explored the large design space of 3D-VRAM arrays and came to a couple of important conclusions that were different from, or not studied in, the conventional 2D cross-point design. We also proposed circuit/architecture optimizations to relax the peripheral overheads of 3D-VRAM and further reduces its cost-per-bit. The system-level evaluations showed that our optimized 3D-VRAM design has better IOPS/$ than other contenders for storage memory in most cases and has the best IOPS/$/J in all tested cases.

References

[1] S. Chung *et al.*, "Fully integrated 54nm stt-ram with the smallest bit cell dimension for high density memory application," in *Proceedings of IEEE International Electron Devices Meeting (IEDM),*, Dec 2010, pp. 12.7.1–12.7.4.

[2] K. Tsuchida *et al.*, "A 64mb mram with clamped-reference and adequate-reference schemes," in *Proceedings of IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),*, Feb 2010, pp. 258–259.

[3] Y. Choi *et al.*, "A 20nm 1.8v 8gb pram with 40mb/s program bandwidth," in *Proceedings of the IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),*, feb. 2012, pp. 46 –48.

[4] B. C. Lee *et al.*, "Architecting phase change memory as a scalable dram alternative," in *Proceedings of the 36th annual international symposium on Computer architecture (ISCA)*. New York, NY, USA: ACM, 2009, pp. 2–13.

[5] H. S. P. Wong *et al.*, "Metal oxide RRAM," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951 –1970, June 2012.

[6] T. yi Liu *et al.*, "A 130.7mm2 2-layer 32gb reram memory device in 24nm technology," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),*, Feb 2013, pp. 210–211.

[7] I. Baek *et al.*, "Realization of vertical resistive memory (VRRAM) using cost effective 3D process," in *Proceedings of the IEEE International Electron Devices Meeting (IEDM)*, 2011, pp. 31.8.1–31.8.4.

[8] W. Chien *et al.*, "Multi-layer sidewall wox resistive memory suitable for 3d reram," in *Proceedings of the IEEE Symposium on VLSI Technology (VLSIT),*, 2012, pp. 153–154.

[9] H.-Y. Chen *et al.*, "HfOx based vertical resistive random access memory for cost-effective 3d cross-point architecture without cell selector," in *Proceedings of the IEEE International Electron Devices Meeting (IEDM),*, 2012, pp. 20.7.1–20.7.4.

[10] C. Chevallier *et al.*, "A 0.13 um 64mb multi-layered conductive metal-oxide memory," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),*, Feb 2010, pp. 260–261.

[11] Y.-C. Chen *et al.*, "3d-him: A 3d high-density interleaved memory for bipolar rram design," in *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH),*, June 2011, pp. 59–64.

[12] A. Kawahara *et al.*, "An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput," in *Proccedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2012, pp. 432–434.

[13] C. Xu *et al.*, "Modeling and design analysis of 3d vertical resistive memory - a low cost cross-point architecture," in *Asia and South Pacific Design Automation Conference (ASP-DAC),*, Jan 2014, pp. 825–830.

[14] C.-W. Hsu *et al.*, "3d vertical taox/tio2 rram with over 103 self-rectifying ratio and sub-ua operating current," in *Proceedings of the IEEE International Electron Devices Meeting (IEDM),*, Dec 2013, pp. 10.4.1–10.4.4.

[15] E. Cha *et al.*, "Nanoscale (10nm) 3d vertical reram and nbo2 threshold selector with tin electrode," in *Proceedings of the IEEE International Electron Devices Meeting (IEDM),*, Dec 2013, pp. 10.5.1–10.5.4.

[16] S. Yu *et al.*, "3d vertical rram - scaling limit analysis and demonstration of 3d array operation," in *Symposium on VLSI Technology (VLSIT),*, June 2013.

[17] J. Liang and H. S. P. Wong, "Cross-point memory array without cell selectors - device characteristics and data storage pattern dependencies," *IEEE Transactions on Electron Devices,*, vol. 57, no. 10, pp. 2531–2538, Oct 2010.

[18] D. Niu *et al.*, "Design trade-offs for high density cross-point resistive memory," in *Proceedings of the ACM/IEEE international symposium on Low power electronics and design (ISLPED)*, 2012, pp. 209–214.

[19] J. Jang *et al.*, "Vertical cell array using tcat (terabit cell array transistor) technology for ultra high density nand flash memory," in *Symposium on VLSI Technology,*, June 2009, pp. 192–193.

[20] K.-T. Park *et al.*, "Three-dimensional 128gb mlc vertical nand flash-memory with 24-wl stacked layers and 50mb/s high-speed programming," in *Proceedings of the IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),*, Feb 2014, pp. 334–335.

[21] N. Agrawal *et al.*, "Design tradeoffs for ssd performance," in *USENIX 2008 Annual Technical Conference on Annual Technical Conference*, ser. ATC'08, 2008, pp. 57–70.

[22] M. Jung, J. Shalf, and K. Mahmut, "Design of a large-scale storage-class rram system," in *Proceedings of the International ACM Conference on International Conference on Supercomputing*, ser. ICS '13, 2013, pp. 103–114.

[23] E. Seevinck, P. van Beers, and H. Ontrop, "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's," *IEEE Journal of Solid-State Circuits,*, vol. 26, no. 4, pp. 525–536, Apr 1991.

[24] M.-F. Chang *et al.*, "A 0.5v 4mb logic-process compatible embedded resistive ram (reram) in 65nm cmos using low-voltage current-mode sensing scheme with 45ns random read time," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),*, Feb 2012, pp. 434–436.

[25] M.-F. Chang *et al.*, "An offset-tolerant current-sampling-based sense amplifier for sub-100na-cell-current nonvolatile memory," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),*, Feb 2011, pp. 206–208.

[26] X. Dong *et al.*, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994–1007, 2012. Available: http://nvsim.org

[27] IC Knowledge LLC., "IC cost model revision 1202a." Available: http://www.icknowledge.com

[28] "SPC TRACE FILE FORMAT SPECIFICATION," *Storage Performance Council*, vol. Tech. Report, no. Rev. 1.0.1, 2002.