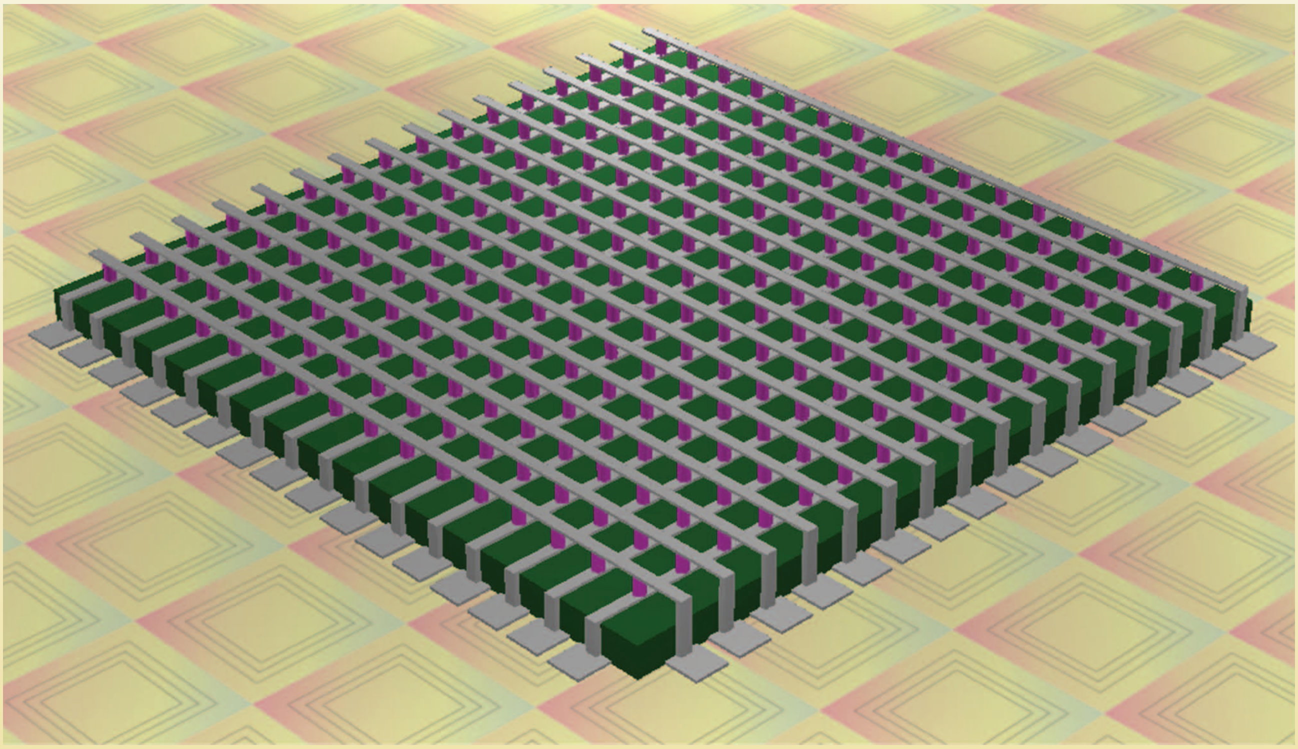
Shimeng Yu and Pai-Yu Chen

Emerging Memory Technologies

Recent Trends and Prospects

This tutorial introduces the basics of emerging nonvolatile memory (NVM) technologies including spin-transfer-torque magnetic random access memory (STT-MRAM), phase-change random access memory (PCRAM), and resistive random access memory (RRAM). Emerging NVM cell characteristics are summarized, and device-level engineering trends are discussed. Emerging NVM array architectures are introduced, including the one-transistor–one-resistor (1T1R) array and the cross-point array with selectors. Design challenges such as scaling the write current and minimizing the sneak path current in cross-point array are analyzed. Recent progress on megabit- to gigabit-level prototype chip demonstrations is summarized. Finally, the prospective applications of emerging NVM are discussed, ranging from the last-level cache to the storage-class memory in the memory hierarchy. Topics of three-dimensional (3D) integration and radiation-hard NVM are discussed. Novel applications beyond the conventional memory applications are also surveyed, including physical unclonable function for hardware security, reconfigurable routing switch for field-programmable gate array (FPGA), logic-in-memory and nonvolatile cache/register/flip-flop

Digital Object Identifier 10.1109/MSSC.2016.2546199
Date of publication: 21 June 2016



1943-0582/16©2016IEEE

for nonvolatile processor, and synaptic device for neuro-inspired computing.

Overview of Emerging Memory Technologies

The functionality and performance of today's computing system are increasingly dependent on the characteristics of the memory subsystem. The memory subsystem has a well-known memory hierarchy: Today static random-access memory (SRAM), dynamic random-access memory (DRAM), and flash are the mainstream memory technologies serving as cache, main memory, and storage memory such as solid-state drive (SSD), respectively. Moving up the memory hierarchy toward the cache, the memory write/read latency decreases. Moving down the memory hierarchy toward the storage, the memory capacity increases. These mainstream memory technologies are essentially based on the charge storage mechanism: SRAM stores the charges at the storage nodes of the cross-coupled inverters, DRAM stores the charges at the cell capacitor, and flash stores the charges at the floating gate of the transistor. All these charge-based memories face challenges in scaling down to the 10-nm node and beyond. The easy loss of the stored charges at nanoscale results in the degradation of performance, reliability, and noise margin. In this context, emerging memory technologies that are noncharge based are actively under research and development in the industry, with the hope of revolutionizing the memory hierarchy [1].

The ideal characteristics for a memory device include fast write/read speed (<ns), low operation voltage (<1 V), low energy consumption (~fJ/b for write/read), long data retention time (>10 years), long write/read cycling endurance (>10^{17} cycles), and excellent scalability (<10 nm). Nevertheless, it is almost impossible to satisfy all of these ideal characteristics in a single "universal" memory device. Several resistance-based emerging NVM technologies have been pursued toward achieving part of these ideal characteristics. The emerging NVM candidates include STT-MRAM [2], PCRAM [3], and RRAM [4].

These emerging NVM technologies share some common features: they are nonvolatile two-terminal devices, and they differentiate their states by the switching between a high resistance state (HRS, or off state) and a low resistance state (LRS, or on state). The switching from off state to on state is called "set," and the switching from on state to off state is called "reset." The transition between the two states can be triggered by an electrical stimulus (i.e., voltage or current pulse). However, the detailed switching physics is quite different: STT-MRAM relies on the parallel configuration (corresponding to LRS) and antiparallel configuration (corresponding to HRS) of two ferromagnetic layers separated by a thin tunneling insulator layer; PCRAM relies on chalcogenide materials to switch between the crystalline phase (corresponding to LRS) and the amorphous phase (corresponding to HRS); and RRAM relies on the formation (corresponding to LRS) and the rupture (corresponding to HRS) of conductive filaments in the insulator between two electrodes. Table 1 compares the typical device characteristics of the emerging memory technologies and the mainstream memory technologies.

Due to the different underlying physics, the device characteristics are also different among emerging NVMs. Therefore, different emerging NVMs may have different application spaces due to their unique characteristics. As compared to SRAM, STT-MRAM has an advantage of a smaller cell area, while STT-MRAM has maintained low programming voltage, fast write/read speed, and long endurance. Thus, STT-MRAM is attractive as a replacement for embedded

TABLE 1. DEVICE CHARACTERISTICS OF MAINSTREAM AND EMERGING MEMORY TECHNOLOGIES.

	MAINSTREAM MEMORIES				EMERGING MEMORIES		
			FLASH				
	SRAM	DRAM	NOR	NAND	STT-MRAM	PCRAM	RRAM
Cell area	>100 F^2	6 F^2	10 F^2	<4F^2 (3D)	6~50F^2	4~30F^2	4~12F^2
Multibit	1	1	2	3	1	2	2
Voltage	<1 V	<1 V	>10 V	>10 V	<1.5 V	<3 V	<3 V
Read time	~1 ns	~10 ns	~50 ns	~10 µs	<10 ns	<10 ns	<10 ns
Write time	~1 ns	~10 ns	10 µs–1 ms	100 µs–1 ms	<10 ns	~50 ns	<10 ns
Retention	N/A	~64 ms	>10 y	>10 y	>10 y	>10 y	>10 y
Endurance	>1E16	>1E16	>1E5	>1E4	>1E15	>1E9	>1E6~1E12
Write energy (J/bit)	~fJ	~10fJ	~100pJ	~10fJ	~0.1pJ	~10pJ	~0.1 pJ

Notes: F: feature size of the lithography. The energy estimation is on the cell-level (not on the array-level). PCRAM and RRAM can achieve less than 4F^2 through 3D integration. The numbers of this table are representative (not the best or the worst cases).

memories (e.g., SRAM or embedded DRAM) in the last-level cache [5]. As compared to flash, PCRAM/RRAM is attractive due to its lower programming voltage and faster write/read speed. Thus, the PCRAM/RRAM is attractive as a replacement for NOR flash for code storage and, more ambitiously, to replace NAND flash for data storage [6]. Besides replacing the existing technologies, the emerging NVM technologies hold the potential to revolutionize today's memory hierarchy by adding more levels in the hierarchy, e.g., creating a storage-class memory layer between the main memory and the SSD [7]. In addition, hybrid systems with emerging memories and mainstream memories are also attractive, e.g., using RRAM as the cache for NAND flash [8].

Emerging NVM Cell Structures and Device-Level /Engineering Challenges

Despite the aforementioned attractive features, emerging NVM technologies face challenges from aspects of process compatibility, manufacturing yield, performance variability, and reliability. In the following, we will discuss the challenges and recent trends of each NVM candidate at the device level.

STT-MRAM Cells

STT-MRAM is based on the magnetic tunnel junction (MTJ) structure. The tunneling magnetoresistance (TMR) ratio (defined as $R_{ap}/R_p - 1$) of the MTJ is typically small (<200% or <2X), thereby imposing challenges for sensing circuit design to sense the small difference between the on and off states. It is also well known that a tradeoff exists between the thermal stability (E_a/kT) and critical write current density (J_c) due to an energy barrier between the parallel and antiparallel states of the MTJ. Given the application demands, the data retention requirement may be relaxed to reduce the write power, e.g., for the last-level cache in which the data are frequently updated. The current trend of STT-MRAM cell design is to switch from the in-plane MTJ [Figure 1(a)] to the perpendicular MTJ [Figure 1(b)] to allow better scalability, longer retention, and lower J_c [9], [10]. The in-plane MTJ's scalability is limited by the aspect ratio (length/width in the lateral dimension) of the cell, as a sufficient shape anisotropy is required for thermal stability, while

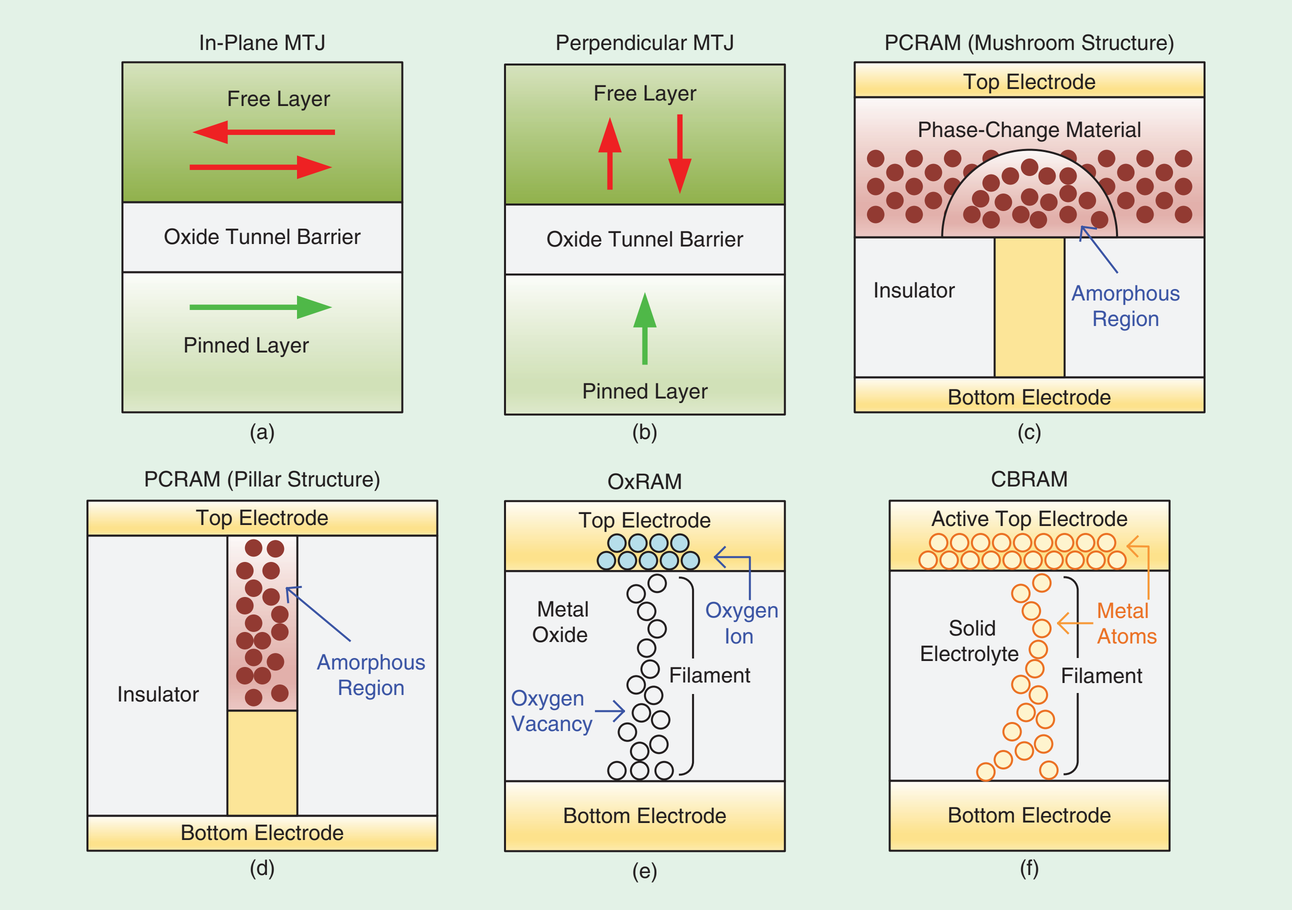


FIGURE 1: A schematic of emerging NVM device structures. (a) STT-MRAM with in-plane MTJ structure. (b) STT-MRAM with perpendicular MTJ structure, which allows better scalability. (c) PCRAM with mushroom structure. (d) PCRAM with thermally confined pillar structure for reducing write current. (e) RRAM based on oxygen vacancies in the filament in the oxide, referred to as OxRAM. (f) RRAM based on metal ions diffusion from active electrode to form conductive bridge in solid electrolyte, referred to as CBRAM.

the perpendicular MTJ achieves the shape anisotropy in the vertical dimension, allowing better scalability in the lateral dimension. State-of-the-art perpendicular MTJ cells have been scaled to 15 nm [11], [12].

Today STT-MRAM's process and manufacturing technology are relatively mature. However, STT-MRAM has relatively poor process compatibility with mainstream silicon CMOS technology, because more than ten layers of exotic ferromagnetic materials (typically CoFeB/MgO) are used in the MTJ stack. In addition, keeping within the thermal budget to crystallize all the magnetic layers while maintaining downward CMOS doping profiles is challenging. The precise deposition and etching to avoid the formation of dead layers/regions of the complicated MTJ stack add another significant cost barrier for foundries to widely adopt this technology.

The functionality and performance of today's computing system are increasingly dependent on the characteristics of the memory subsystem.

PCRAM Cells

The PCRAM cells are typically based on GST materials (e.g., $Ge_2Sb_2Te_5$). The GST material systems can further be tuned for device characteristics that are of interest. For example, Ge-rich GST (*N*-doped) could be used to achieve better data retention for high temperature automotive applications [13]. The PCRAM's on/off resistance ratio is much larger (in the range of 100–1,000×) than STT-MRAM. Thus, in principle, multilevel cell (MLC) operations are allowed (even 4 b/cell are feasible [14]). The key challenge for PCRAM cell design is the relatively large write current required to melt the phase-change materials. Even for state-of-the-art PCRAM at 20 nm, the write current (~100 μA [15]) is roughly three to ten times larger than its STT-MRAM or RRAM counterparts. The current trend of PCRAM cell design is to switch from the mushroom cell [Figure 1(c)] to the pillar cell [Figure 1(d)] to confine the heat dissipation, thereby reducing the write current. An extremely scaled PCRAM cell using the carbon-tube electrodes suggests that the write current can achieve ~1 μA at 2-nm node [16]. The PCRAM's switching speed (>50 ns) is limited by the slow crystalline process, also ten times longer than its RRAM counterparts, while the PCRAM's endurance (10^6~10^9 cycles) is comparable to that of the RRAM. The PCRAM's data retention (especially for the MLC) is limited by resistance drift due to the relaxation of the amorphous state. Thus, sophisticated circuit-level compensation schemes are needed [17]. Despite the fact that the PCRAM's cell characteristics are less competitive than RRAM in terms of the write power and speed, today PCRAM's process and manufacturing technology is quite mature. PCRAM has generally good process compatibility with mainstream silicon CMOS technology, as GST materials can be deposited by sputtering under back-end-of-line (BEOL) temperature (<400 °C).

RRAM Cells

There are two subcategories within RRAM: oxide-RAM (OxRAM) and conductive bridge RAM (CBRAM). The difference is that OxRAM's filament consists of oxygen vacancies in the oxide layer [Figure 1(e)], while CBRAM's filament consists of metal atoms, formed by fast-diffusive Ag or Cu ions migrating into the solid-electrolyte [Figure 1(f)]. Despite different underlying physics, these two types of RRAMs share many common device characteristics. The only notable difference may be that OxRAM's on/off resistance ratio may be smaller (in the range of 10–100×) and offers better endurance up to 10^{12} cycles, while CBRAM's on/off resistance ratio can be quite large (10^3–10^6×) but with limited endurance (<10^4 cycles) [18]. The switching of RRAM includes unipolar and bipolar modes depending on the oxide and electrode materials system [19]. The unipolar mode generally requires larger write current and shows less endurance; thus, the bipolar mode is preferred. The key challenge of RRAM cell design is the variability of the switching parameters. Owing to the stochastic nature of ionic (oxygen vacancies or metal ions) migration, the filament shape varies from device to device and also from cycle to cycle (within one device). Remarkable variation in resistance distribution (which can be one or two orders of magnitude) adds challenges to the sensing circuit design and requires the write-verify techniques to program to the target states, which could be latency consuming for the MLC operations. Although RRAM could require smaller write current (e.g., ~10 μA) due to the filamentary switching mechanism, the data retention may be problematic when filament is too thin [20], and at the same time the random telegraph noise due to the filament instability may become significant [21]. Nevertheless, the scalability to 2-nm node of RRAM cell has been demonstrated by sidewall electrodes [22]. RRAM has generally excellent process compatibility with the mainstream silicon CMOS technology, as many RRAM materials (e.g., HfO_x, TaO_x) are already used in silicon transistors' high-*k* dielectric process. Atomic-layer deposition allows for the accurate deposition of RRAM thin film under BEOL temperature (<400 °C).

Emerging NVM Array Architectures and Circuit-Level Design Challenges

1T1R Array Architecture

One of the common emerging NVM array architectures is the 1T1R array.

In this design, each NVM cell is in series with a cell selection transistor, as shown in Figure 2(a). The addition of a selection transistor is able to isolate the selected cell from other unselected cells. The word line (WL) controls the gate of the transistor; thus, tuning the WL voltage can control the write current that is delivered to the NVM cell. The NVM cell's top electrode connects to the bit line (BL) while its bottom electrode connects to the contact via to the drain of the transistor. The source line (SL) connects to the source of the transistor. The typical cell area of 1T1R array is 12 F^2 (F is the lithography feature size) if the gate width/length (W/L) of the transistor is one. The minimum cell area can be reduced to 6 F^2 if the aggressive borderless DRAM design rule with sharing BL and SL is applied. The cell area will be increased if the W/L of the transistor is larger than one when a minimum-sized transistor cannot provide sufficient write current (usually in cases of STT-MRAM and PCRAM).

Figure 2(a)–(c) shows the typical write/read scheme for the 1T1R array. For the set operation, WL voltage is applied to turn on the transistor of the selected cell, and a write voltage is applied to the BL of the selected cell while SL is grounded; for the reset operation, WL voltage is applied to turn on the selection transistor of the selected cell, and a write voltage is applied to the SL of the selected cell while BL is grounded to reverse the current, as the typical STT-MRAM and RRAM operation needs "bipolar" switching (PCRAM does not need to reverse the voltage polarity though). For unselected rows and columns, the WL, BL, and SL are all grounded. To read out the data from the 1T1R array, WL voltage is applied to turn on the selection transistor of the selected cell, and a read voltage is applied to the BL while SL is grounded. The sense amplifier (S/A) thus can sense the difference in the read-out current for HRS and LRS through the BL with a reference. Because the transistors are off for the unselected cells, there are no crosstalk or interference issues in the 1T1R array, and each cell can be independently and randomly accessed. Multiple bits can be written (or read) in parallel into (or from) the same row by activating multiple columns.

The 1T1R array faces scaling challenges if the NVM's write current cannot scale accordingly. Figure 2(d) shows the silicon CMOS low-power logic transistor's drive current with the scaling from 130 nm down to 20 nm for different W/L simulated with the Predictive Technology Model (PTM) [23]. The representative write current of STT-MRAM, PCRAM, and RRAM from the literature data is also marked. RRAM's write current typically ranges from 10 to 100 µA (with some scattered data points for sub-10 µA), and it does not scale with the device area due to the filamentary conduction mechanism. Although STT-MRAM and PCRAM's write current scales with the device area, at 20-nm node, STT-MRAM's current is ~40 µA, and PCRAM's current is ~100 µA. It is seen that in most cases, $W/L = 1$ transistors could not provide sufficient write current for a NVM cell; thus, 6 F^2 cell area is unlikely to achieve using logic-compatible process. Although the transistor's gate voltage (V_{gs}) can be boosted under the specially engineered DRAM process to increase the drive current, it is still limited. For example, the drive current at V_{gs} = 5 V can approach 40 µA for a $W/L = 1$ transistor at 27-nm node [24]. Therefore, reducing the write current down to sub-10 µA by device engineering is of great importance for continuing the scaling of the 1T1R array. In addition, reducing the write voltage down to sub-1 V is also necessary for embedded applications if using a logic-compatible process.

Cross-Point Array Architecture

The other common array architecture is the cross-point (X-point, or crossbar) array, which consists of

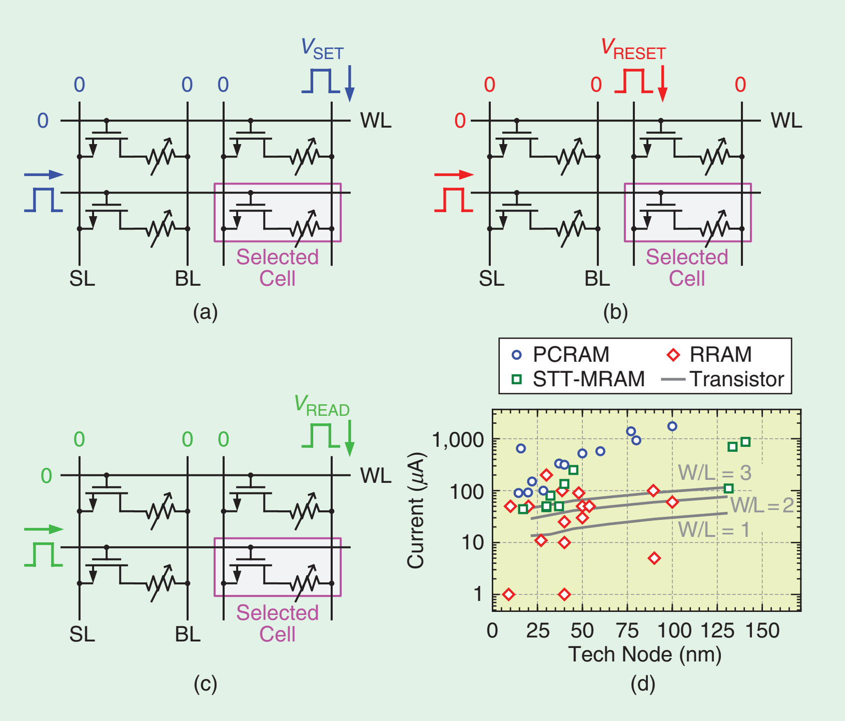


FIGURE 2: A schematic of the 1T1R array: (a) set scheme, (b) reset scheme, and (c) read scheme. (d) A silicon CMOS low-power logic transistor's drive current with the scaling from 130 nm down to 20 nm for different W/L simulated with the PTM model [23]. The representative write current of STT-MRAM, PCRAM, and RRAM from the literature data is also marked.

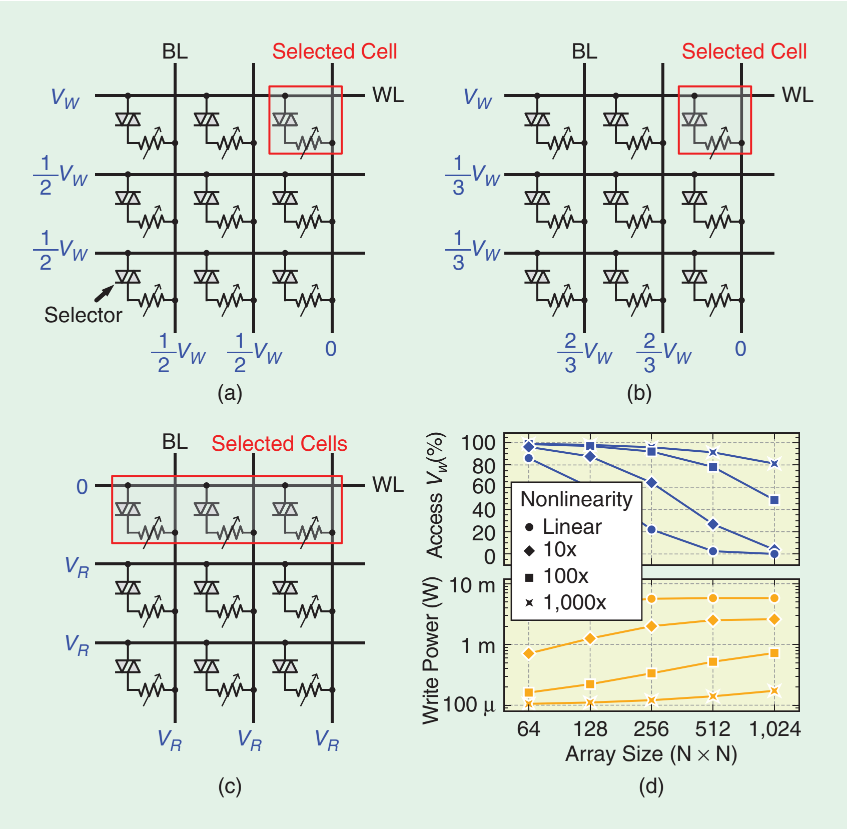


FIGURE 3: A schematic of the cross-point array. The selector is added in series with the RRAM cell at each cross-point. (a) V/2 write scheme. (b) V/3 write scheme. (c) Read scheme (for an entire row). (d) SPICE simulation of write margin and write power as a function of cross-point array size. Increasing the NVM cell's I-V nonlinearity (*N*) by adding selectors is helpful to minimize the IR drop problem along the interconnect wire and the sneak path problem.

rows and columns perpendicular to each other with NVM cells sandwiched in between, as shown in Figure 3(a). The cross-point array, in principle, can achieve a 4 F^2 cell area; thus, it can achieve higher integration density than the 1T1R array. Typically, the selectors are added in series with the NVM cells to prevent cross talk or interference between cells in the cross-point array, which is referred to as one-selector and one-resistor (1S1R) architecture. The cross-point array can support PCRAM and RRAM but generally does not support STT-MRAM because of a very small on/off ratio (~2×); thus, the sense margin becomes indistinguishable due to the sneak path current.

The write/read schemes of the cross-point array are as follows. To successfully program the NVM cells, two common write schemes (V/2 and V/3) can be applied. Figure 3(a) shows the voltage bias conditions for the V/2 scheme. In the V/2 scheme, for the set operation, the selected cell's WL and BL are biased at the write voltage V_w and ground, respectively. For the reset operation, the bias conditions on WL and BL are reversed for the bipolar switching. In both set and reset operations, all the unselected WLs and BLs are biased at $V_w/2$. Therefore, only the selected cell sees a full V_w, while the half-selected cells along the selected WL or BL see a half V_w and all the other unselected cells in the array see zero voltage [in reality, due to the current-resistance (IR) drop along the interconnect, the voltage is not perfectly zero though]. Here the assumption is that $V_w/2$ should not disturb the half-selected cell's resistance. Figure 3(b) shows the voltage bias conditions for the V/3 scheme. In the V/3 scheme, for the set operation, the selected cell's WL and BL are biased at the write voltage V_w and ground, respectively. For the reset operation, the bias conditions on WL and BL are reversed for the bipolar switching. The unselected WLs and BLs are biased at 1/3 V_w and 2/3 V_w for the set operation, respectively. The unselected WLs and BLs are biased at 2/3 V_w and 1/3 V_w for the reset operation, respectively. In this way, the selected cell sees V_w, while all other unselected cells in the array only see 1/3 V_w. Here the assumption relaxes so that 1/3 V_w should not disturb the unselected cell's resistance. The pros and cons of these two write schemes can be summarized as follows: the V/2 scheme typically has less power consumption than the V/3 scheme. This is because the unselected cells (not along the selected WL and BL) in the V/2 scheme see zero voltage ideally, while all the unselected cells in the V/3 scheme see 1/3 V_w, thus consuming static power during the write period. On the other hand, the V/3 scheme has better immunity to the write disturbance than the V/2 scheme, as the maximum voltage that the unselected cells see is 1/3 V_w in the V/3 scheme while is 1/2 V_w in the V/2 scheme. It is possible to have multiple-bit parallel write in the cross-point array with either the V/2 or V/3 scheme by biasing multiple BLs (or WLs) to be ground in the set (or reset) operation. The penalty for multiple-bit parallel write is a larger driver size at each row (or column) as it has to deliver the multiple write current in addition to the sneak path via the unselected cells. Figure 3(c) shows the read scheme for the cross-point array. All the columns are biased at the read voltage V_r, while the selected row is biased at ground and the unselected rows are biased at V_r. Therefore, only the cells of the selected row see a read voltage and all the other unselected cells see zero voltage (in reality, due to the IR drop along interconnect, the voltage is not perfectly zero though). The entire selected row can be read-out in parallel if each column can have one S/A. However, the pitch of S/As is typically larger than the column pitch; thus, multiple columns have to share one S/A. S/As can be generally categorized into two types [25]: voltage mode and current mode. In practical designs, the choice between voltage-mode sensing and current-mode sensing depends on the array size and the NVM cell characteristics. The general conclusion is that for an array with a long BL length or a higher LRS resistance (smaller read-out current), current sensing provides faster access.

The cross-point array suffers from two well-known design challenges: 1) the IR drop problem along the interconnect wire and 2) the sneak path problem through the unselected cells. The IR drop problem becomes significant when the WL and BL wire width scales to sub-50-nm regime where the interconnect resistivity drastically increases due to the increased electron surface scattering. For example, at 20-nm node, the copper interconnect resistance between two neighboring cells is ~2.93 Ω; thus, the IR drop along the wire for a large array (e.g., a 1,024 × 1,024 array) is no longer negligible. The farthest cell from the driver sees an interconnect resistance ~3 kΩ. If the NVM cell's resistance (typically a few kΩ up to tens of kΩ) is comparable to this interconnect resistance, a portion of the write voltage will drop on the wire instead of the NVM cell. To guarantee a successful write operation, the write voltage provided from the driver has to be boosted over the actual switching voltage of the NVM cell to compensate for the IR drop. However, the write voltage cannot be boosted too much because 1/2 V_w (in the V/2 scheme) should not disturb the NVM resistance for the cells close to the driver.

The sneak path problem is associated with the IR drop problem. Take the V/2 scheme as an example. The half-selected cells along the selected WL and BL form the sneak paths during the write operation. The sneak paths contribute additional current to the IR drop and further degrade the write margin. Meanwhile, the sneak paths increase the write current (thus the write power) that is provided by the driver transistors at the edge of the cross-point array. Further discussions about the IR drop problem and the sneak path problem of the cross-point array architecture can be found in [26]–[28]. The conclusions from these works indicate that increasing the LRS resistance (or equivalently reducing the write current) and increasing the I–V nonlinearity of the NVM cell (with the help of the selector) are useful to minimize the IR drop and sneak paths. Figure 3(d) shows the SPICE simulation of the write margin and write power as a function of the cross-point array size for different I–V nonlinearity (N). The nonlinearity is defined as the current ratio between V_w and $V_w/2$. The NVM cell resistance is fixed at 40 kΩ, and the wire width is fixed at 20 nm in this study. It is seen that at least $N > 1{,}000$ is needed for maintaining sufficient write margin and minimizing write power for a large array (e.g., a 1,024 × 1,024 array).

Selector Device for Cross-Point Array

Next we will survey the two-terminal selector devices reported in [29]. For unipolar switching (e.g., PCRAM), a p-n diode is the most common device for the cell selector. Although a high-performance p-n diode is easily fabricated with epitaxial silicon technology for the planar device structure, it is not feasible to implement an epitaxial silicon-based p-n diode at the BEOL for 3D integration because it is difficult to grow epitaxial silicon on a metal layer and a high processing temperature is required. On the other hand, amorphous silicon allows for a BEOL processing temperature (<400 °C). But an amorphous silicon p-n diode does not meet the requirement for the current density for the NVM programming. For bipolar switching (e.g., RRAM), bidirectional nonlinearity is required (note: a bidirectional selector also works for unipolar PCRAM).

There are two types of bidirectional selectors: Type I: exponential I–V and Type II: threshold I–V. Figure 4(a) and (b) shows the representative I–V

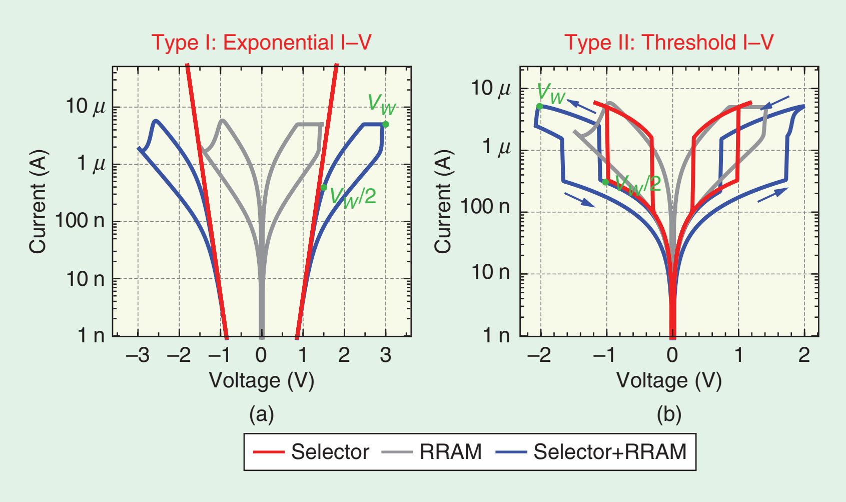


FIGURE 4: Representative I-V characteristics for a bipolar RRAM with (a) Type I selector and a bipolar RRAM with (b) Type II selector. SPICE simulation is performed with the RRAM compact model [38].

characteristics for a bipolar RRAM with Type I selector and a bipolar RRAM with Type II selector, respectively. The metrics of selector performance are 1) the nonlinearity (N) defined as the current ratio between V_w and $V_w/2$, which will determine how effective the sneak current suppression is, and 2) the drive current density (e.g., 10 MA/cm^2 is required for 10-μA write current at the 10-nm node). Oxide/electrode interface engineering or oxide/oxide bandgap engineering with the tunneling current mechanism can be leveraged as Type I selector, e.g., Ni/TiO$_2$/Ni bidirectional selector [30]. In addition, Cu ion motion in the Cu-containing mixed-ionic-electronic-conduction materials also show a good bidirectional exponential I–V for bipolar switching RRAM, as demonstrated by a series of works [31]–[33]. The aforementioned Type I selectors rely on an exponential slope in the I–V curve to turn on the selector, accompanied with an increase of the current by several orders of magnitude. Ideally, an abrupt turn-on behavior with steep slope is preferred, which is referred to as the threshold switching (Type II). The threshold selector typically exhibits a hysteresis in I–V as it turns on above a threshold voltage and turns off below a hold voltage. Threshold switching can be achieved in the metal-insulator-transition (MIT) Mott oxide materials such as NbO$_2$ [34]. The drawback of MIT-based threshold selectors is a relatively small nonlinearity (typically $N < 100$). Besides Mott oxide materials, an ovonic-threshold-switch based on chalcogenide materials has been demonstrated to be an excellent threshold selector [35]. Recently, a field-assisted-superlinear-threshold selector has been reported [36], which shows outstanding nonlinearity ($N > 10^7$), steep turn-on slope (<5 mV/dec) and high current drivability (>5 MA/cm^2), and the threshold voltage is claimed to be adjustable from ~0.3 V to ~1.3 V to match various RRAM characteristics.

Although substantial progress has been made in the past few years, the development of selector devices is still a key challenge for implementing large-scale cross-point memory architectures today. Most importantly, the selector device characteristics must match the NVM device characteristics. Adding selector devices in series with the NVM cell inevitably increases the programming voltage as part of the voltage is used to turn on the selector device. For the bidirectional selector with exponential I–V, the read sense margin generally degrades because the read-current for LRS is also suppressed. As a result, it requires a much longer time for sensing. As a reference, a well-designed current-mode sense amplifier can sense sub-100 nA read-out current within 26 ns [37]. For the threshold switching selector with abrupt I–V, the read voltage has to be boosted above the threshold voltage of the selector; thus, it runs a risk of disturbing the NVM resistance in the read operation. Ultimately, it is preferred that the NVM cell itself has a built-in I–V nonlinearity thereby eliminating the necessity of the external selector device.

Merging NVM technologies face challenges from aspects of process compatibility, manufacturing yield, performance variability, and reliability.

Recent Progress on Prototype Chip Demonstration

There are two methods for integrating RRAM cells on top of the CMOS circuits. The first approach is to fabricate the RRAM cells following the front-end-of-line process (close to the transistor fabrication at a lower-level interconnect). For example, the RRAM cells can be deposited at the contact via between the drain and metal 1, and this approach is typically employed in the 1T1R array architecture. The second approach is to fabricate the RRAM cells via the BEOL process at the top-level interconnect (decoupled from the transistor fabrication). For example, the RRAM cells can be deposited at the contact via between metal 4 and metal 5. One of the advantages of the BEOL integration is that the peripheral circuits can be hidden underneath the cross-point array to save the area as demonstrated in Panasonic's 8-Mb prototype chip [39].

Figure 5 summarizes the recent prototype chip demonstrations of various NVM technologies reported in the major conferences. Figure 5(a) shows the memory capacity versus year, and Figure 5(b) shows the write/read bandwidth versus year. PCRAM and RRAM have demonstrated >Gb-level capacity owing to the smaller cell size (4 F^2 using cross-point array or 6 F^2 using 1T1R array with minimum sized transistor), while STT-MRAM's capacity is only up to the 64-Mb level (cell size is still >30 F^2 owing to a larger transistor used to deliver sufficient write current in the 1T1R array and a relaxed layout design rule). It is noted that the bandwidth is related to the input/output (I/O) interface. NAND flash typically use page-program (e.g., 4 kb per page) to achieve high bandwidth, although it has slow write time per cell. Emerging NVM macros typically do not use wide-I/O (only 64- or 128-b interface) but has fast write-time per cell. Despite the narrow I/O, the emerging NVMs remarkably improve the write/read bandwidth over the NAND or NOR flash. The data sheet of the prototype chip parameters (e.g., capacity, performance, etc.) can be accessed via the Arizona State University Memory Trend [40].

Next we will present a few representative prototypes for each emerging NVM.

STT-MRAM Prototypes

Toshiba reported a 64-Mb STT-MRAM prototype chip using a 65-nm CMOS

Different emerging NVM devices may have different application spaces in the memory hierarchy due to their unique characteristics.

process, featuring a 30-ns cycle time [41]; more recently, Toshiba reported a 1-Mb STT-MRAM prototype chip using a 65-nm CMOS process, improving the write/read cycle to 3 ns [42]. TSMC reported a 1-Mb STT-MRAM prototype chip using a 40-nm CMOS process and featuring a 10-ns read cycle time [43]. Qualcomm/TDK-Headway reported a 1-Mb STT-MRAM prototype chip using a 40-nm CMOS process, featuring a 20-ns write/read cycle time [44].

PCRAM Prototypes

Numonyx (now Micron) reported a 1-Gb 1T1R PCRAM prototype [45]. The cell selection device is a bipolar junction transistor. The fabrication was done in 45-nm process, and 9-MB/s write bandwidth and 266-MB/s read bandwidth were demonstrated. Samsung reported an 8-Gb cross-point PCRAM prototype chip [46]. The cell selection device is a silicon p-n diode. The fabrication was done in a 20-nm process. A 40-MB/s write bandwidth has been demonstrated.

RRAM Prototypes

ITRI reported a 4-Mb 1T1R HfO_x-based RRAM prototype chip [47]. The fabrication was done in a 180-nm CMOS process. A single-level-cell operation with 7.2-ns read/write random access was presented, and an MLC 2 b/cell operation with 160-ns write-verify scheme was demonstrated. Panasonic reported an 8-Mb cross-point TaO_x-based RRAM prototype chip [39]. The fabrication was done in a 180-nm CMOS process. A 443-MB/s write throughput (64-b parallel write per 17.2-ns cycle) and a 25-ns read access has been demonstrated. For the embedded applications, National Tsing-Hua University reported a 4-Mb macro in 65-nm logic-compatible process [48] and a 1-Mb macro in a 28-nm logic-compatible process [49]. For the large-capacity standalone applications, SanDisk/Toshiba reported a 32-Gb cross-point OxRAM prototype chip with a 24-nm cell half-pitch [50]. Both Panasonic and SanDisk/Toshiba's design adopted a two-layer stacked cross-point array architecture by sharing the BL to increase the integration density (similar to Intel/Micron's 3D X-point architecture [51]). Recently, Micron/Sony reported a 16-Gb 1T1R CBRAM prototype chip [52]. The fabrication was done in a 27-nm DRAM-like process. A 200-MB/s write bandwidth and 1-GB/s read bandwidth have been demonstrated.

Applications of Emerging NVM

Replacing Existing Technologies in Memory Hierarchy

Different emerging NVM devices may have different application spaces in the memory hierarchy due

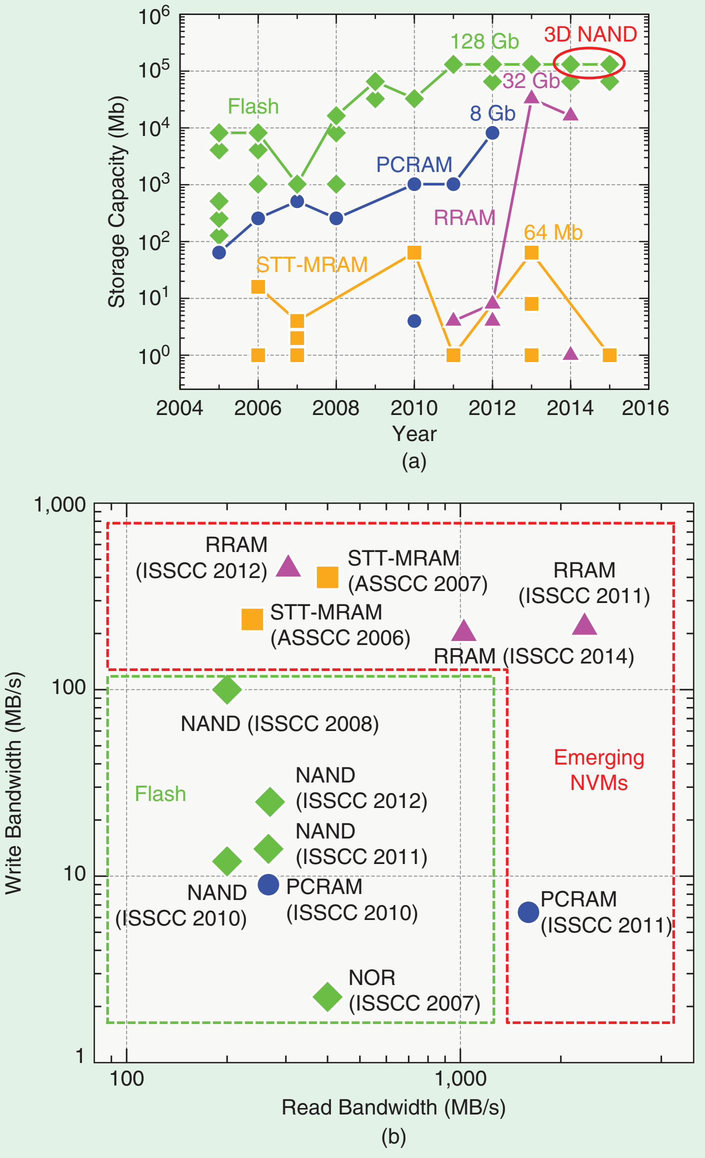


FIGURE 5: A summary of (a) memory capacity versus years and (b) write/read bandwidth versus years. The data are collected from major conferences such as IEDM, ISSCC, and VLSI from 2005 to 2015 [40].

to their unique characteristics. STT-MRAM has advantages of fast write/read speed (<10 ns) and long endurance ($>10^{15}$ cycles) that PCRAM/RRAM lacks; thus, STT-MRAM is attractive to replace SRAM or embedded DRAM in the last-level cache. Today's STT-MRAM has >30 F^2 cell area; thus it is not economically competitive to replace DRAM for main memory. In the long run, with the reduction of write current of the STT-MRAM, it may have the potential to achieve 6 F^2 cell area. Then it may become attractive to replace DRAM for main memory because it does not need the periodic refresh. On the other hand, PCRAM/RRAM has advantages of a smaller cell area (4 F^2~6 F^2) and, thus, a larger capacity than STT-MRAM; however, its endurance is typically in the range of 10^6–10^{10} cycles; thus it is more suitable for the storage-class memory filling the gap between the main memory and the storage memory (e.g., SSD). With the smart architectural wear-leveling techniques [53], the requirement of the cycling endurance for the main memory system can be reduced to 10^{10} cycles; PCRAM/RRAM may meet this requirement with the help of a strong error-correct code. It is not economically promising for PCRAM/RRAM to directly replace the NAND flash for standalone SSD due to a higher cost per bit. State-of-the-art two-dimensional (2D) NAND flash has been scaled down to around 15 nm in 2015, while the 3D stackable NAND flash is emerging [54], [55]. Twenty-four-, 32-, and 48-layer (up to 256 Gb) 3D NAND flash chips featuring MLC have been demonstrated [56]–[58]. On the market, 3D NAND flash-based SSD has been commercialized. Although at the single device level, PCRAM/RRAM outperforms NAND flash in many aspects such as faster programming speed, smaller programming voltage, and better endurance. The key challenge for PCRAM/RRAM to compete with NAND flash is the integration density or, more importantly, the cost per bit. To approach similar device density as the 3D NAND flash, a technology path toward the 3D stackable PCRAM/RRAM is required. Ultimately, the 3D stacked PCRAM/RRAM may have potential to serve as high-end enterprise SSD if cost per bit is not of the highest priority but performance is.

Ultimately, the 3D stacked PCRAM/RRAM may have potential to serve as high-end enterprise SSD if cost per bit is not of highest priority but performance is.

3D Integration of PCRAM/RRAM

There are two monolithic 3D integration approaches for PCRAM and/or RRAM technologies: one is based on stacking the conventional horizontal cross-point array layer by layer [59], as shown in Figure 6(a), which is referred to as 3D X-point as in the Intel/Micron announcement [51]; the other is the vertical pillar structure with RRAM sandwiched between the pillar electrodes and multilayer plane electrodes [60], as shown in Figure 6(b). This vertical 3D RRAM concept is similar to today's vertical channel 3D NAND flash. Figure 6(c) shows the cross-section microscopic schematic of the vertical RRAM prototype cell [61] by cutting through one pillar electrode: the RRAM cells are formed at the sidewall of the pillar electrode and in contact with the plane electrodes (highlighted by the red-dash circle), and there is one cell at each metal layer. The fabrication cost of the first approach using the stacked horizontal cross-point array is relatively higher because the number of lithography steps increases with the number of the layers; thus, the fabrication cost remains high as the lithography step and mask are expensive.

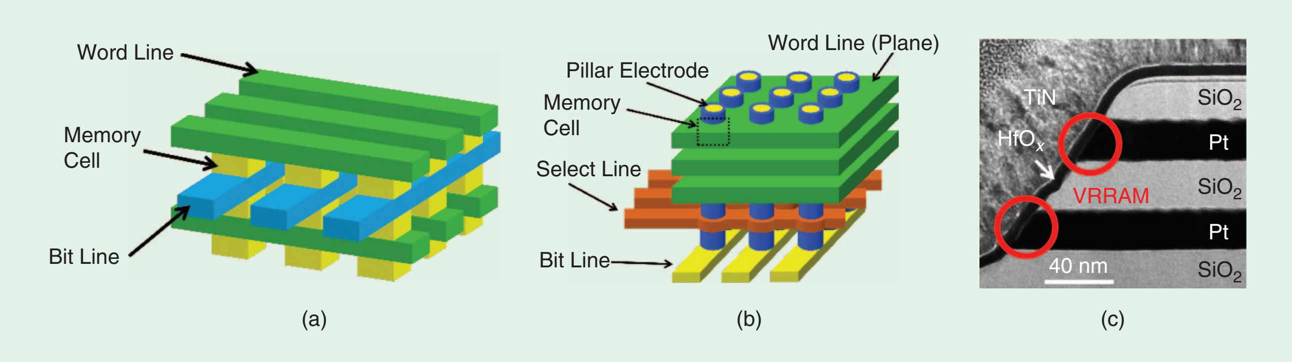


FIGURE 6: The schematics of (a) a stacked 3D horizontal cross-point array and (b) a 3D vertical cross-point array. (c) The microscopic cross section of a 3D vertical RRAM prototype by cutting through one pillar electrode. Adapted from [61].

The second approach using vertical RRAM requires only one critical lithography step to define the pillar electrodes after sequentially depositing multilayer plane electrodes, making it a more promising approach for reducing the fabrication cost. However, the cost per bit analysis for these two 3D array architectures is not sp intuitive. Although the vertical RRAM saves the fabrication cost, its minimal F is not as small as that of the horizontal counterpart; thus, it has a lower integration density because the diameter of the pillar electrode is limited by the following factors. First, the aspect ratio of the pillar electrode is limited by the etching process capability of metal/dielectric multilayers. Second, the pillar electrode resistance will drastically increase at the nanoscale. As a rough estimation, the vertical RRAM can scale to F=30 nm considering a pillar diameter (~20 nm) plus twice of the RRAM oxide thickness (~5 nm). If the horizontal RRAM can scale to F=10 nm with the help of the selector, then one layer of horizontal RRAM has the same integration density as nine layers of vertical RRAM.

A further detailed analysis is needed to assess the pros and cons of these two 3D integration approaches. The 3D horizontal cross-point array still needs the help of the selectors to address the sneak path problem as in the case of 2D cross-point array discussed earlier. The 3D vertical cross-point array prefers the built-in I–V nonlinearity of the RRAM cell as it is difficult to add the external selector on the sidewall. The problem is that the middle electrode between the selector and the RRAM cell will make a short circuit of multiple layers.

Niche Market of Emerging NVMs

Emerging NVMs may find applications in a niche market. One example is as radiation-hard NVM for aerospace electronics. Many experiments show that RRAM is robust against the radiation effects such as total ionizing dose effect [62], while the single-event-upset effect observed in the RRAM 1T1R array was attributed to the photocurrent generated at the neighboring transistor's drain to body p-n junction [63], which can be eliminated by using silicon-on-insulator transistors. Besides the standalone NVM applications, emerging NVMs are also suitable for embedded applications. RRAM devices are especially attractive due to their good compatibility with logic processes. For instance, an embedded RRAM solution has been introduced for a 28-nm technology node [64], [65]. Therefore, RRAM has great potential as MB-level embedded NVM for micro-controller applications.

Beyond the conventional memory applications, novel applications that use emerging NVM are arising.

Novel Applications of Emerging NVMs

Beyond the conventional memory applications, novel applications that use emerging NVM are arising. For instance, the use of emerging NVM as the physical unclonable function as hardware security primitive has been proposed [66], [67], which leverages the intrinsic variations in the emerging NVM's switching processes. The use of RRAM as the reconfigurable switch has also been proposed. RRAM-based FPGA was designed [68] and demonstrated [69]. The use of RRAM as ternary content-addressable-memory for fast-searching big data has been reported [70]. Adding RRAM cells on top of the SRAM cell enables the instant-on and instant-off power gating by storing data from SRAM to RRAM locally before going to standby mode, which saves the latency and power to transfer the data to off-chip or embedded flash, as shown in Figure 7. Example of nonvolatile cache include eight-transistor–two-resistor (8T2R) nonvolatile SRAM [71] and the seven-transistor–one-resistor (7T1R) nonvolatile SRAM [72]. The same principle also applies to the nonvolatile register and nonvolatile flip-flop design.

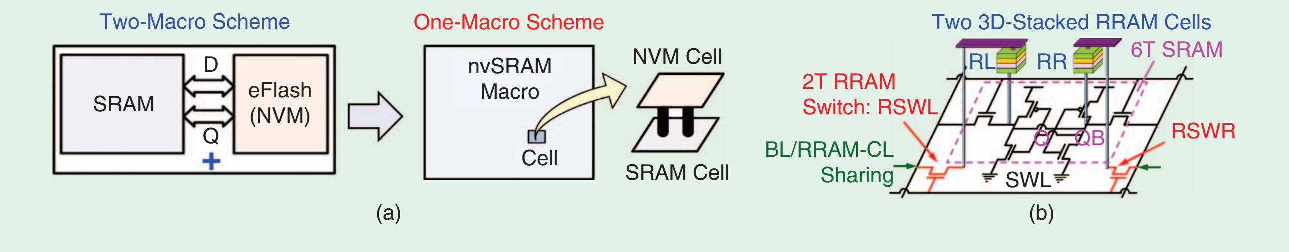


FIGURE 7: (a) Power-gating schemes: (a) two-macro with conventional embedded flash and one-macro with embedded NVM cells on top SRAM cells. (b) Eight-transistor-two-resistor (8T2R) nonvolatile SRAM. Adapted from [71].

Owing to the long endurance of STT-MRAM, it can be hybrid with logic gates (e.g., full adder or multiplier) to enable fine-grain power gating [73] and realize the logic-in-memory and eventually the nonvolatile processor. Another emerging application is using PCRAM/RRAM as synaptic devices for hardware implementation of neuro-inspired computing [74]. Owing to PCRAM/RRAM's multilevel capability, it serves as analog memory emulating the function of plastic synapses in a neural network, and the cross-point array architecture can efficiently implement the weighted sum and weight update process in the learning algorithms in an analog computing fashion [75].

Outlook

Although the early vision for emerging NVM is to replace an existing mainstream memory technology in the memory hierarchy, it is not entirely clear that these goals continue to make sense, given the many diverse potential applications of emerging NVM. By taking advantage of emerging NVM, there are enormous opportunities to completely rethink the design of the computer system to gain orders of magnitude improvement in speed and/or power consumption. Emerging NVM's unique physical properties may also add new functionality and features to the systems. A revolution of the future's computing paradigm will bring about a fundamental change in how one can extract benefits out of the technology advancements.

Acknowledgments

This work is in part supported by NSF CCF-1449653. The authors thank Dr. Nuo Xu of Samsung Semiconductor, Prof. Meng-Fan Chang of National Tsinghua University, and Dr. Tim Hollis of Micron Technology for the discussions and suggestions on this article.

References

[1] H. S. P. Wong, and S. Salahuddin, "Memory leads the way to better computing," *Nat. Nanotechnol.*, vol. 10, pp. 191–194, Mar. 2015.

[2] B. J. Zhu, "Magnetoresistive random access memory: the path to competitiveness and scalability," *Proc. IEEE*, vol. 96, no. 11, pp. 1786–1798, 2008.

[3] H. S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory," *Proc. IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[4] H. S. P. Wong, H. Y. Lee, S. Yu, Y. S. Chen, Y. Wu, P. S. Chen, B. Lee, F. T. Chen, and M. J. Tsai, "Metal–oxide RRAM," *Proc. IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012.

[5] S. P. Park, S. Gupta, N. Mojumder, A. Raghunathan, and K. Roy, "Future cache design using STT MRAMs for improved energy efficiency: devices, circuits and architecture," in *Proc. 49th ACM Design Automation Conf.,* San Francisco, CA, 2012, pp. 492–497.

[6] M. Jung, J. Shalf, and M. Kandemir, "Design of a large-scale storage-class RRAM system," in *Proc. ACM Int. Conf. Supercomputing*, 2013, pp. 103–114.

[7] R. F. Freitas, and W. W. Wilcke, "Storage-class memory: The next storage system technology," *IBM J. Res. Develop.*, vol. 52, no. 4.5, pp. 439–447, 2008.

[8] S. Tanakamaru, H. Yamazawa, T. Tokutomi, S. Ning, and K. Takeuchi, "Hybrid storage of ReRAM/TLC NAND Flash with RAID-5/6 for cloud data centers," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, 2014, pp. 336–337.

[9] T. Kishi, H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama, T. Daibou, M. Nagamine, M. Amano, S. Takahashi, M. Nakayama, N. Shimomura, H. Aikawa, S. Ikegawa, S. Yuasa, K. Yakushiji, H. Kubota, A. Fukushima, and M. Oogane, "Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM," in *Proc. IEEE Int. Electron Devices Meeting*, San Francisco, CA, 2008, pp. 1–4.

[10] L. Thomas, G. Jan, J. Zhu, H. Liu, Y. J. Lee, S. Le, R. Y. Tong, K. Pi, Y. J. Wang, D. Shen, R. He, J. Haq, J. Teng, V. Lam, K. Huang, T. Zhong, T. Torng, and P. K. Wang, "Perpendicular spin transfer torque magnetic random access memories with high spin torque efficiency and thermal stability for embedded applications," *J. Appl. Phys.*, vol. 115, no. 17, pp. 172615, 2014.

[11] J. H. Kim, W. C. Lim, U. H. Pi, J. M. Lee, W. K. Kim, J. H. Kim, K. W. Kim, Y. S. Park, S. H. Park, M. A. Kang, Y. H. Kim, W. J. Kim, S. Y. Kim, J. H. Park, S. C. Lee, Y. J. Lee, J. M. Yoon, S. C. Oh, S. O. Park, S. Jeong, S. W. Nam, H. K. Kang, and E. S. Jung, "Verification on the extreme scalability of STT-MRAM without loss of thermal stability below 15 nm MTJ cell," in *Proc. IEEE Symp. VLSI Technology*, Honolulu, HI, 2014, pp. 1–2.

[12] N. Xu, J. Wang, Y. Lu, H.-H. Park, B. Fu, R. Chen, W. Choi, D. Apalkov, S. Lee, S. Ahn, Y. Kim, Y. Nishizawa, K.-H. Lee, Y. Park, and E. S. Jung, "Physics-based compact modeling framework for state-of-the-art and emerging STT-MRAM technology," in *Proc. IEEE Int. Electron Devices Meeting,* Washington, DC, 2015, pp. 28.5.1–28.5.4.

[13] H. Y. Cheng, J. Y. Wu, R. Cheek, S. Raoux, M. BrightSky, D. Garbin, S. Kim, T. H. Hsu, Y. Zhu, E. K. Lai, E. Joseph, A. Schrott, S. C. Lai, A. Ray, H. L. Lung, and C. Lam, "A thermally robust phase change memory by engineering the Ge/N concentration in (Ge, N)xSbyTez phase change material," in *Proc. IEEE Int. Electron Devices Meeting,* San Francisco, CA, 2012, pp. 31.1.1–31.1.4.

[14] T. Nirschl, J.B. Phipp, T.D. Happ, G.W. Burr, B. Rajendran, M.-H. Lee, A. Schrott, M. Yang, M. Breitwisch, C.-F. Chen, E. Joseph, M. Lamorey, R. Cheek, S.-H. Chen, S. Zaidi, S. Raoux, Y.C. Chen, Y. Zhu, R. Bergmann, H.-L. Lung, and C. Lam, "Write strategies for 2 and 4-bit multi-level phase-change memory," in *Proc. IEEE Int. Electron Devices Meeting,* Washington, DC, 2007, pp. 461–464.

[15] M. J. Kang, T. J. Park, Y. W. Kwon, D. H. Ahn, Y. S. Kang, H. Jeong, S. J. Ahn, Y. J. Song, B. C. Kim, S.W. Nam, H. K. Kang, G. T. Jeong, and C. H. Chung, "PRAM cell technology and characterization in 20nm node size," in *Proc. IEEE Int. Electron Devices Meeting,* Washington, DC, 2011, pp. 3.1.1–3.1.4.

[16] J. Liang, R. G. D. Jeyasingh, H.-Y. Chen, and H.-S. P. Wong, "A 1.4μA reset current phase change memory cell with integrated carbon nanotube electrodes for cross-point memory application," in *Proc. IEEE Symp. VLSI Technology*, Honolulu, HI, 2011, pp. 100–101.

[17] W.-S. Khwa, M.-F. Chang, J.-Y. Wu, M.-H. Lee, T.-H. Su, K.-H. Yang, T.-F. Chen, T.-Y. Wang, H.-P. Li, M. BrightSky, S. Kim, H.-L. Lung, and C. Lam, "A resistance-drift compensation scheme to reduce MLC PCM raw BER by Over 100X for storage class memory applications," in *Proc. IEEE Int. Solid-State Circuit Conf.*, 2016, pp. 134–136.

[18] E. Vianello, O. Thomas, G. Molas, O. Turkyilmaz, N. Jovanovic, D. Garbin, G. Palma, M. Alayan, C. Nguyen, J. Coignus, B. Giraud, T. Benoist, M. Reyboz, A. Toffoli, C. Charpin, F. Clermidy, and L. Perniola, "Resistive memories for ultra-low-power embedded computing design," in *Proc. IEEE Int. Electron Devices Meeting,* San Francisco, CA, 2014, pp. 6.3.1–6.3.4.

[19] S. Yu and H. S. P. Wong, "A phenomenological model for the reset mechanism of metal oxide RRAM," *IEEE Electron Device Lett.*, vol. 31, no. 12, pp. 1455–1457, 2010.

[20] Y. Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini, N. Raghavan, S. Clima, L. Zhang, A. Belmonte, A. Redolfi, G. S. Kar, G. Groeseneken, D. J. Wouters, and M. Jurczak, "Improvement of data retention in HfO2/Hf 1T1R RRAM cell under low operating current," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, 2013, pp. 10.1.1–10.1.4.

[21] S Ambrogio, S Balatti, A Cubeta, A Calderoni, N Ramaswamy, and D Ielmini, "Understanding switching variability and random telegraph noise in resistive RAM," in *Proc. IEEE Int. Electron Devices Meeting,* Washington, DC, 2013, pp. 31.5.1–31.5.4.

[22] K.-S. Li, C. H. Ho, M.-T. Lee, M.-C. Chen, C.-L. Hsu, J. M. Lu, C. H. Lin, C. C. Chen, B. W. Wu, Y. F. Hou, C. Y. Lin, Y. J. Chen, T. Y. Lai, M. Y. Li, I. Yang, C. S. Wu, and F.-L. Yang, "Utilizing Sub-5 nm sidewall electrode technology for atomic-scale resistive memory fabrication," in *Proc. IEEE Symp. VLSI Technology,* Honolulu, HI, 2014, pp. 1–2.

[23] Arizona State University. (2011, June 1). Predictive Technology Model (PTM). [Online]. Available: http://ptm.asu.edu/

[24] J. Zahurak, K. Miyata, M. Fischer, M. Balakrishnan, S. Chhajed, D. Wells, H. Li, A. Torsi, J. Lim, and M. Korber, "Process integration of a 27nm, 16Gb Cu ReRAM," in *Proc. IEEE Int. Electron Devices Meeting*, San Francisco, CA, 2014, pp. 6.2.1–6.2.4.

[25] M. F. Chang, A. Lee, P. C. Chen, C. J. Lin, Y. C. King, S. S. Sheu, and T. K. Ku, "Challenges and circuit techniques for energy-efficient on-chip nonvolatile memory using memristive devices," *IEEE J. Emerging Selected Topics Circuits Syst.*, vol. 5, no. 2, pp. 183–193, 2015.

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

[27] Y. Deng, P. Huang, B. Chen, X. Yang, B. Gao, J. Wang, L. Zeng, G. Du, J. Kang, and X. Liu, "ReRAM crossbar array with cell selection device: A device and circuit interaction study," *IEEE Trans. Electron Devices*, vol. 60, no. 2, pp. 719–726, 2013.

[28] D. Niu, C. Xu, N. Muralimanohar, N. P. Jouppi, and Y. Xie, "Design trade-offs for high density cross-point resistive memory," in *Proc. ACM/IEEE Int. Symp. Low Power Electronics and Design,* 2012, pp. 209–214.

[29] G. W. Burr, R. S. Shenoy, K. Virwani, P. Narayanan, A. Padilla, B. Kurdi, and H. Hwang, "Access devices for 3D crosspoint memory," *J. Vac. Sci. Technol. B, Micorelectron. Process. Phenom.*, vol. 32, no. 4, pp. 40802, 2014.

[30] J.-J. Huang, Y.-M. Tseng, W.-C. Luo, C.-W. Hsu, and T.-H. Hou, "One selector-one resistor (1S1R) crossbar array for high-density flexible memory applications," in *Proc. IEEE Int. Electron Devices Meeting*, 2011, pp. 733–736.

[31] K. Gopalakrishnan, R. S. Shenoy, C. T. Rettner, K. Virwani, D. S. Bethune, R. M. Shelby, G. W. Burr, A. Kellock, R. S. King, K. Nguyen, A. N. Bowers, M. Jurich, B. Jackson, A. M. Friz, T. Topuria, P. M. Rice, and B. N. Kurdi, "Highly-scalable novel access device based on mixed ionic electronic conduction (MIEC) materials for high density phase change memory (PCM) arrays," in *Proc. IEEE Symp. VLSI Technology,* Honolulu, HI, 2010, pp. 205–206.

[32] G. W. Burr, K. Virwani, R. S. Shenoy, G. Fraczak, C. T. Rettner, A. Padilla, R. S. King, K. Nguyen, A. N. Bowers, M. Jurich, M. BrightSky, E. A. Joseph, A. J. Kellock, N. Arellano, B. N. Kurdi, and K. Gopalakrishnan, "Recovery dynamics and fast (sub-50ns) read operation with access devices for 3D crosspoint memory based on mixed-ionic-electronic-conduction (MIEC)," in *Proc. IEEE Symp.VLSI Technology*, Kyoto, Japan, 2013, pp. T66–T67.

[33] K. Virwani, G. W. Burr, R. S. Shenoy, C. T. Rettner, A. Padilla, T. Topuria, P. M. Rice, G. Ho, R. S. King, K. Nguyen, A. N. Bowers, M. Jurich, M. BrightSky, E. A. Joseph, A. J. Kellock, N. Arellano, B. N. Kurdi, and K. Gopalakrishnan, "Sub-30nm scaling and high-speed operation of fully-confined access-devices for 3D crosspoint memory based on mixed-ionic-electronic-conduction (MIEC) materials," in *Proc. IEEE Int. Electron Devices Meeting*, San Francisco, CA, 2012, pp. 2.7.1–2.7.4.

[34] S.G. Kim, T.J. Ha, S. Kim, J.Y. Lee, K.W. Kim, J.H. Shin, Y.T. Park, S.P. Song, B.Y. Kim, W.G. Kim, J.C. Lee, H.S. Lee, J.H. Song, E.R. Hwang, S.H. Cho, J.C. Ku, J.I. Kim, K.S. Kim, J. H. Yoo, H.J. Kim, H.G. Jung, K.J. Lee, S. Chung, J.H. Kang, J. H. Lee, H. S. Kim, S. J. Hong, G. Gibson, and Y. Jeon, "Improvement of characteristics of NbO2 selector and full integration of 4F2 2x-nm tech 1S1R ReRAM," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, 2015, pp. 10.3.1–10.3.4.

[35] D. Kau, S. Tang, I. V. Karpov, R. Dodge, B. Klehn, J. Kalb, J. Strand, A. Diaz, N. Leung, J. Wu, and S. Lee, "A stackable cross point phase change memory," in *Proc. IEEE Int. Electron Devices Meeting,* Baltimore, MD, 2009, pp. 1–4.

[36] S. H. Jo, T. Kumar, S. Narayanan, W. D. Lu, and H. Nazarian, "3D-stackable crossbar resistive memory based on field assisted superlinear threshold (FAST) selector," in *Proc. IEEE Int. Electron Devices Meeting*, San Francisco, CA, 2014, pp. 6.7.1–6.7.4.

[37] M. F. Chang, S. J. Shen, C. C. Liu, C. W. Wu, Y. F. Lin, Y. C. King, C. J. Lin, H. J. Liao, Y. D. Chih, and H. Yamauchi, "An offset-tolerant fast-random-read current-sampling-based sense amplifier for small-cell-current nonvolatile memory," *IEEE J. Solid-State Circuits*, vol. 48, no. 3, pp. 864–877, 2013.

[38] P. Y. Chen, and S. Yu, "Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design," *IEEE Trans. Electron Devices*, vol. 62, no. 12, pp. 4022–4028, 2015.

[39] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, K. Tanabe, T. Nakamura, Y. Sumimoto, N. Yamada, N. Nakai, S. Sakamoto, Y. Hayakawa, K. Tsuji, S. Yoneda, A. Himeno, K. Origasa, K. Shimakawa, T. Takagi, T. Mikawa, and K. Aono, "An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2012pp. 178–185.

[40] Arizona State University. (2015, Apr. 3). ASU Memory Chip Trend. [Online]. Available: http://faculty.engineering.asu.edu/shimengyu/model-downloads/.

[41] K. Tsuchida, T. Inaba, K. Fujita, Y. Ueda, T. Shimizu, Y. Asao, T. Kajiyama, M. Iwayama, K. Sugiura, S. Ikegawa, T. Kishi, T. Kai, M. Amano, N. Shimomura, H. Yoda, and Y. Watanabe, "A 64Mb MRAM with clamped-reference and adequate-reference schemes," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, 2010, pp. 258–259.

[42] H. Noguchi, K. Ikegami, K. Kushida, K. Abe, S. Itai, S. Takaya, N. Shimomura, J. Ito, A. Kawasumi, H. Hara, and S. Fujita, "A 3.3ns-access-time 71.2μW/MHz 1Mb embedded STT-MRAM using physically eliminated read-disturb scheme and normally-off memory architecture," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, 2015, pp. 1–3.

[43] H.-C. Yu, K.-C. Lin, K.-F. Lin, C.-Y. Huang, Y.-D. Chih, T.-C. Ong, J. Chang, S. Natarajan, and L. Tran, "Cycling endurance optimization scheme for 1Mb STT-MRAM in 40nm technology," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, 2013, pp. 224–225.

[44] L. Yu, T. Zhong, W. Hsu, S. Kim, X. Lu, J. J. Kan, C. Park, W. C. Chen, X. Li, X. Zhu, P. Wang, M. Gottwald, J. Fatehi, L. Seward, J. P. Kim, N. Yu, G. Jan, J. Haq, S. Le, Y. J. Wang, L. Thomas, J. Zhu, H. Liu, Y. J. Lee, R. Y. Tong, K. Pi, D. Shen, R. He, Z. Teng, V. Lam, R. Annapragada, T. Torng, P. K. Wang, and S. H. Kang, "Fully functional perpendicular STT-MRAM macro embedded in 40 nm logic for energy-efficient IOT applications," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, 2015, pp. 26.1.1–26.1.4.

[45] C. Villa, D. Mills, G. Barkley, H. Giduturi, S. Schippers, and D. Vimercati, "A 45nm 1Gb 1.8V phase-change memory," *in Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, 2010, pp. 270–271.

[46] Y. Choi, I. Song, M.-H. Park, H. Chung, S. Chang, B. Cho, J. Kim, Y. Oh, D. Kwon, J. Sunwoo, J. Shin, Y. Rho, C. Lee, M. Kang, J. Lee, Y. Kwon, S. Kim, J. Kim, Y.-J. Lee, Q. Wang, S. Cha, S. Ahn, H. Horii, J. Lee, K. Kim, H. Joo, K. Lee, Y.-T. Lee, J. Yoo, and G. Jeong, "A 20nm 1.8V 8Gb PRAM with 40MB/s program bandwidth," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, 2012, pp. 46–48.

[47] S. S. Sheu, M. F. Chang, K. F. Lin, C. W. Wu, Y. S. Chen, P. F. Chiu, C. C. Kuo, Y. S. Yang, P. C. Chiang, W. P. Lin, C. H. Lin, H. Y. Lee, P. Y. Gu, S. M. Wang, F. T. Chen, K. L. Su, C. H. Lien, K. H. Cheng, H. T. Wu, T. K. Ku, M. J. Kao, and M. J. Tsai, "A 4Mb embedded SLC resistive-RAM macro with 7.2ns read-write random-access time and 160ns MLC-access capability," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, 2011, pp. 200–202.

[48] M. F. Chang, C. W. Wu, C. C. Kuo, S. J. Shen, K. F. Lin, S. M. Yang, Y. C. King, C. J. Lin, and Y. D. Chih, "A 0.5V 4Mb logic-process compatible embedded resistive RAM (ReRAM) in 65 nm CMOS using low-voltage current-mode sensing scheme with 45ns random read time," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, 2012, pp. 434–436.

[49] M. F. Chang, J. J. Wu, T. F. Chien, Y. C. Liu, T. C. Yang, W. C. Shen, Y. C. King, C. J. Lin, K. F. Lin, Y. D. Chih, S. Natarajan, and J. Chang, "Embedded 1Mb ReRAM in 28 nm CMOS with 0.27-to-1V read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, 2014, pp. 332–333.

[50] T., Y. Liu, T. H. Yan, R. Scheuerlein, Y. Chen, J. K. Lee, G. Balakrishnan, G. Yee, H. Zhang, A. Yap, J. Ouyang, T. Sasaki, S. Addepalli, A. Al-Shamma, C. Y. Chen, M. Gupta, G. Hilton, S. Joshi, A. Kathuria, V. Lai, D. Masiwal, an M. Matsumoto, "A 130.7 mm 2 2-layer 32Gb ReRAM memory device in 24nm technology," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, 2013, pp. 210–211.

[51] Intel and Micron Produce Breakthrough Memory Technology. (2015). [Online]. Available: http://newsroom.intel.com/community/intel_newsroom/blog/2015/07/28/intel-and-micron-produce-breakthrough-memory-technology

[52] R. Fackenthal, M. Kitagawa, W. Otsuka, K. Prall, D. Mills, K. Tsutsui, J. Javanifard, K. Tedrow, T. Tsushima, Y. Shibahara, and G. Hush, "A 16Gb ReRAM with 200MB/s write and 1GB/s read in 27nm technology," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, 2014, pp. 338–339.

[53] M. K. Qureshi, M. Franchescini, V. Srinivasan, L. Lastras, B. Abali, and J. Karidis, "Enhancing lifetime and security of PCM-based main memory with start-gap wear leveling," in *Proc. IEEE/ACM Int. Symp. Microarchitecture*, 2009, pp. 14–23.

[54] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, "Bit cost scalable technology with punch and plug process for ultra high density flash memory," in

Proc. IEEE Symp. VLSI Technology, Kyoto, Japan, 2007, pp. 14–15.

[55] J. Jang, H. S. Kim, W. Cho, H. Cho, J. Kim, S. I. Shim, Y. Jang, J. H. Jeong, B. K. Son, D. W. Kim, K. Kim, J. J. Shim, J. S. Lim, K. H. Kim, S. Y. Yi, J. Y. Lim, D. Chung, H. C. Moon, S. Hwang, J. W. Lee, Y. H. Son, U. I. Chung, and W. S. Lee, "Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultra high density NAND flash memory," in *Proc. IEEE Symp. VLSI Technology*, Honolulu, HI, 2009, pp. 192–193.

[56] K. T. Park, J. Han, D. Kim, S. Nam, K. Choi, M. S. Kim, P. Kwak, D. Lee, Y. H. Choi, K. M. Kang, M. H. Choi, D. H. Kwak, H. Park, S. Shim, H. J. Yoon, D. Kim, S. Park, K. Lee, K. Ko, D. K. Shim, Y. L. Ahn, J. Park, J. Ryu, D. Kim, K. Yun, J. Kwon, S. Shin, D. S. Byeon, K. Choi, J. M. Han, K. H. Kyung, J. H. Choi, and K. Kim, "Three-dimensional 128Gb MLC vertical NAND Flash memory with 24-WL stacked layers and 50MB/s high-speed programming," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, 2014, pp. 334–335.

[57] J. W. Im, W. P. Jeong, D. H. Kim, S. W. Nam, D. K. Shim, M. H. Choi, H. J. Yoon, D. H. Kim, Y. S. Kim, H. W. Park, D. H. Kwak, S. W. Park, S. M. Yoon, W. G. Hahn, J. H. Ryu, S. W. Shim, K. T. Kang, S. H. Choi, J. D. Ihm, Y. S. Min, and I. M. Kim, "A 128Gb 3b/cell V-NAND flash memory with 1Gb/s I/O rate," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb. 2015, pp. 1–3.

[58] D. Kang, W. Jeong, C. Kim, D.-H. Kim, Y. S. Cho, K.-T. Kang, J. Ryu, K.-M. Kang, S. Lee, W. Kim, H. Lee, J. Yu, N. Choi, D.-S. Jang, J.-D. Ihm, D. Kim, Y.-S. Min, M.-S. Kim, A.-S. Park, J.-I. Son, M. Kim, P. Kwak, B.-K. Jung, D.-S. Lee, H. Kim, H.-J. Yang, D. S. Byeon, K. T. Park, K. H. Kyung, and J. H. Choi, "256Gb 3b/Cell V-NAND flash memory with 48 stacked WL layers," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, 2016, pp. 130–131.

[59] M.-J. Lee, Y. Park, B.-S. Kang, S.-E. Ahn, C. Lee, K. Kim, W. Xianyu, G. Stefanovich, J.-H. Lee, S.-J. Chung, Y.-H. Kim, C.-S. Lee, J.-B. Park, I.-G. Baek, and I.-K. Yoo, "2-stack 1D-1R cross-point structure with oxide diodes as switch elements for high density resistance RAM applications," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, Dec. 2007, pp. 771–774.

[60] I. G. Baek, C. J. Park, H. Ju, D. J. Seong, H. S. Ahn, J. H. Kim, M. K. Yang, S. H. Song, E. M. Kim, S. O. Park, C. H. Park, C. W. Song, G. T. Jeong, S. Choi, H. K. Kang, and C. Chung, "Realization of vertical resistive memory (VRRAM) using cost effective 3D process," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, Dec. 2011 pp. 31.8.1–31.8.4.

[61] H.-Y. Chen, S. Yu, B. Gao, P. Huang, J. F. Kang, and H.-S. P. Wong, "HfOx based vertical RRAM for cost-effective 3D cross-point architecture without cell selector," in *Proc. IEEE Int. Electron Devices Meeting*, San Francisco, CA, Dec. 2012, pp. 20.7.1–20.7.4.

[62] R. Fang, Y. Gonzalez-Velo, W. Chen, K. Holbert, M. Kozicki, H. Barnaby, and S. Yu, "Total ionizing dose effect of γ-ray radiation on the switching characteristics and filament stability of HfOx resistive random access memory," *Appl. Phys. Lett.*, vol. 104, no. 18, pp. 183507, May 7, 2014.

[63] W. G. Bennett, N. C. Hooten, R. D. Schrimpf, R. A. Reed, M. H. Mendenhall, M. L. Alles, J. Bi, E. X. Zhang, D. Linten, M. Jurzak, and A. Fantini, "Single- and multiple-event induced upsets in HfO2/Hf 1T1R RRAM," *IEEE Trans. Nucl. Sci.*, vol. 61, no. 4, pp. 1717–1725, 2014.

[64] Y. Hayakawa, A. Himeno, R. Yasuhara, W. Boullart, E. Vecchio, T. Vandeweyer, T. Witters, D. Crotti, M. Jurczak, S. Fujii, S. Ito, Y. Kawashima, Y. Ikeda, A. Kawahara, K. Kawai, Z. Wei, S. Muraoka, K. Shimakawa, T. Mikawa, and S. Yoneda, "Highly reliable TaOx ReRAM with centralized filament for 28-nm embedded application," in *Proc. IEEE Symp. VLSI Technology,* Kyoto, Japan, June 2015, pp. T14–T15.

[65] M. F. Chang, J. J. Wu, T. F. Chien, Y. C. Liu, T. C. Yang, W. C. Shen, Y. C. King, C. J. Lin, K. F. Lin, Y. D. Chih, S. Natarajan, and J. Chang, "19.4 Embedded 1Mb ReRAM in 28nm CMOS with 0.27-to-1V read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, Feb. 2014, pp. 332–333.

[66] R. Liu, H. Wu, Y. Pang, H. Qian, and S. Yu, "Experimental characterization of physical unclonable function based on 1kb resistive random access memory arrays," *IEEE Electron Device Lett.*, vol. 36, no. 12, pp. 1380–1383, 2015.

[67] A. Chen, "Utilizing the variability of resistive random access memory to implement reconfigurable physical unclonable functions," *IEEE Electron Device Lett.*, vol. 36, no. 2, pp. 138–140, 2015.

[68] S. Tanachutiwat, M. Liu, and W. Wang, "FPGA based on integration of CMOS and RRAM," *IEEE Trans. VLSI Syst.*, vol. 19, no. 11, pp. 2023–2032, 2011.

[69] Y. Y. Liauw, Z. Zhang, W. Kim, A. El Gamal, and S. S. Wong, "Nonvolatile 3D-FPGA with monolithically stacked RRAM-based configuration memory," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, Feb. 2012, pp. 406–408.

[70] M. F. Chang, C. C. Lin, A. Lee, C. C. Kuo, G. H. Yang, H. J. Tsai, T. F. Chen, S. S. Sheu, P. L. Tseng, H. Y. Lee, and T. K. Ku, "17.5 A 3T1R nonvolatile TCAM using MLC ReRAM with Sub-1ns search time," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, Feb. 2015, pp. 1–3.

[71] P.-F. Chiu, M.-F. Chang, S.-S. Sheu, K.-F. Lin, P.-C. Chiang, C.-W. Wu, W.-P. Lin, C.-H. Lin, C.-C. Hsu, F. T. Chen, K.-L. Su, M.-J. Kao, and M.-J. Tsai, "A low store energy, low VDDmin, nonvolatile 8T2R SRAM with 3D stacked RRAM devices for low power mobile applications," in *Proc. IEEE Symp. VLSI Circuits,* Honolulu, HI, June 2010, pp. 229–230.

[72] A. Lee, M.-F. Chang, C.-C. Lin, C.-F. Chen, M.-S. Ho, C.-C. Kuo, P.-L. Tseng, S.-S. Sheu, and T.-K. Ku, "RRAM-based 7T1R nonvolatile SRAM with 2x reduction in store energy and 94x reduction in restore energy for frequent-off instant-on applications," in *Proc. IEEE Symp. VLSI Circuits,* Kyoto, Japan. June 2015, pp. C76–C77.

[73] M. Natsui, D. Suzuki, N. Sakimura, R. Nebashi, Y. Tsuji, A. Morioka, T. Sugibayashi, S. Miura, H. Honjo, K. Kinoshita, and S. Ikeda, "Nonvolatile logic-in-memory array processor in 90nm MTJ/MOS achieving 75% leakage reduction using cycle-based power gating," in *Proc. IEEE Int. Solid-State Circuits Conf.,* San Francisco, CA, Feb. 2013, pp. 194–195.

[74] D. Kuzum, S. Yu, and H. S. P. Wong, "Synaptic electronics: materials, devices and applications," *Nanotechnology*, vol. 24, no. 38, pp. 382001, Sept. 2, 2013.

[75] S. Yu, P.-Y. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *Proc. IEEE Int. Electron Devices Meeting,* Washington, DC, Dec. 2015, pp. 17.3.1–17.3.4.

About the Authors

Shimeng Yu (shimengy@asu.edu) received the B.S. degree in microelectronics from Peking University, Beijing, China, in 2009 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, California, in 2011 and 2013, respectively. He held internships with IMEC, Belgium, in 2011 and the IBM T.J. Watson Research Center in 2012. He is currently an assistant professor of electrical engineering and computer engineering at Arizona State University, Tempe. His research interests are emerging nanodevices and circuits beyond CMOS technology with a focus on the resistive switching memories and new computing paradigms beyond von Neumann architecture with a focus on the brain-inspired neuromorphic computing. He has published more than 50 journal papers and more than 80 conference papers with a citation greater than 2,700 and H-index 25 according to Google Scholar. Among his honors, he was awarded the Stanford Graduate Fellowship from 2009 to 2012, the IEEE Electron Devices Society Masters Student Fellowship in 2010, the IEEE Electron Devices Society Ph.D. Student Fellowship in 2012, the DoD-DTRA Young Investigator Award in 2015, and the NSF CAREER Award in 2016. He is a member of the IEEE Circuits and Systems Society Technical Committee of Nanoelectronics and Gigascale Systems. He is a Member of the IEEE.

Pai-Yu Chen received the B.S. degree in electrical engineering from National Taiwan University, Taipei, in 2010 and the M.S.E. degree in electrical engineering from the University of Texas, Austin, in 2013. He is currently working toward the Ph.D. degree in electrical engineering at Arizona State University, Tempe. His research interest involves emerging nonvolatile memory device/circuit/architecture design and new computing paradigm exploration. Among his honors, he was awarded the Taiwanese government scholarships to study abroad in 2015. He is a Student Member of the IEEE.