Evolution of Phase-Change Memory for the Storage-Class Memory and Beyond

Taehoon Kim, *Senior Member, IEEE,* and Seungyun Lee

(Invited Paper)

***Abstract*— In this article, the development history and the technical hurdles of phase-change memory (PCM) are reviewed and recent progress and future directions are discussed. Prospects of PCM for storage-class memory (SCM) are discussed in terms of the technical challenges to satisfy the market requirements, mainly for the performance and cost effectiveness. For a more in-depth discussion, PCM is segmented into the memory part, the access device part, and the sensing scheme part. In the memory part, Set (crystallization)–Reset (amorphization) write characteristics and thermal disturbance (TDB) will be reinterpreted in terms of power consumption and performance of SCM. In the access device part, the application history of various devices, such as the transistor, the diode, and the two-terminal selector, will be reviewed and explained based on the process integration issues and area efficiency. In the next part, various sensing schemes used to make a lower read latency, multilevel cell (MLC), and cross point (X-point) are summarized. Thereafter, future directions and possible evolution pathways of 3-D structures, including X-point and vertical X-point (VXP), will be discussed for the first time. Finally, the feasibility of PCM for neuromorphic application will be followed.**

***Index Terms*— Cross-point (X-point) memory, neuromorphic device, phase-change memory (PCM), reset current, resistive random access memory, scaling limit, storage-class memory (SCM), synaptic device, thermal disturbance (TDB).**

I. Introduction

PHASE-CHANGE memory (PCM) has been one of the most promising future memory devices for a long time, along with RRAM, STT-MRAM, and ferroelectric random access memory (FeRAM) [1]. Strictly speaking, PCM is not a future memory but a prototype memory, since it has been produced by a couple of companies, such as Numonyx (now Micron) and Samsung [2], [3]. As a nonvolatile memory (NVM) with low latency and decent reliability, PCM has attracted people's attention as a new potential solution for memory systems, as it is different from the other memory solutions, such as DRAM and NAND Flash memory. The competitiveness of PCM as a memory solution has been further improved with the downscaling of cell size, because it enabled the reduction of Reset current (I_{Reset}) and resulted in less power consumption. Furthermore, the evolution of access devices from large transistors to the vertical diode was a huge breakthrough in reducing the overhead area of core circuits for cell operation, which consequently strengthens the competitiveness of PCM by increasing the cost effectiveness or decreasing process cost per bit [4].

Therefore, the goal of the first-generation PCM was to improve overall performances and cost effectiveness to compete with DRAM: latency for read and write, cyclic write endurance, retention, and so on. Some reports showed very promising results; for example, various reports showed a read latency of 120 ns, a write latency of 150 ns [5], and a write cyclic endurance of more than 1E9 [6]. However, in spite of such great advantages, PCM was not so successful in settling down in the memory market, as DRAM and NAND Flash have also continued to evolve. It seems difficult to catch up with DRAM's performance and NAND's cost effectiveness, especially 3-D NAND Flash (3-D NAND).

However, with the advent of the memory-centric computing era, the necessity of storage-class memory (SCM) has been on the rise [7], since it can fill the performance gap between DRAM and 3-D NAND. Fig. 1(a) shows the SCM's relative position compared with DRAM and 3-D NAND in terms of access latency and cost effectiveness, and Fig. 1(b) shows the position of SCM in the memory hierarchy.

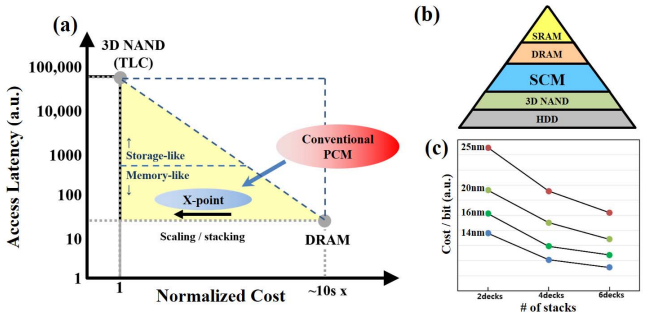

Fig. 1. (a) Suggested position of SCM (X-point PCM) in the market, compared with DRAM and 3-D NAND in terms of access latency and normalized cost (i.e., cost per bit) [7]. (b) Position of SCM in the memory hierarchy based on the access latency [7]. (c) Cost per bit of X-point PCM by lateral scaling and vertical stack-up; assuming that all core circuits are packed under the cell area that is for minimum areal overhead and maximum die per wafer [7].

Manuscript received December 4, 2019; accepted January 2, 2020. Date of publication February 5, 2020; date of current version March 24, 2020. The review of this article was arranged by Editor C. V. Mouli. *(Corresponding author: Taehoon Kim.)*

The authors are with SK Hynix Inc., Gyeonggi 17336, South Korea (e-mail: taehoon12.kim@sk.com).

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

Digital Object Identifier 10.1109/TED.2020.2964640

The position of SCM should be located between DRAM and 3-D NAND, particularly in the lower yellow triangle in Fig. 1(a). This signifies that SCM should be cheaper than DRAM and faster than 3-D NAND. Moving toward the bottom-left region of the yellow triangle will further guarantee the success of the SCM in the memory market. In order for this to happen, the read latency and cost effectiveness of PCM need to be improved further to play a role as an SCM. However, simple downscaling of the conventional PCM cannot achieve these improvements.

Cross-point (X-point) PCM has emerged as a solution for these issues, since it has not only an ultimate areal density of $4F^2$ and 3-D stackability for superior cost effectiveness but also a low read latency. The ideal $4F^2$ architecture and low read latency were possible because of the development of a two-terminal selector, simply referred to as "selector" hereafter. We will further explain this in Sections II-C, III-B, and III-E. The 3-D stackability was possible due to the development of the material technology of both the amorphous material (selector) and the polycrystalline material (memory). Normally, conventional silicon devices, such as MOSFET and p-n diode, should be made on single crystal silicon; however, large area single crystal silicon cannot be grown on the amorphous oxide. The stackability driven cost effectiveness is shown in Fig. 1(c), assuming that all core circuits are packed under the cell area for minimum areal overhead. The cost effectiveness can be increased by adding more stacks, despite the increased total number of process steps. The calculation for cost effectiveness considered all costs of processes, such as pitch double, pitch quadruple, and ArF-i photo. In spite of these huge benefits, only a few research reports were made on the technology of X-point PCM [8], [9] and no report was made for the future direction of the technology.

In this article, we will make a review of the overall technology of PCM from the conventional type to the X-point type for SCM application. For the conventional PCM, the basic theories, device operations, and characteristics of materials are well documented along with the issues and hurdles, which have been discussed extensively [10]–[12]. On the other hand, only a few reports on X-point PCM have been published [8], [9]. Therefore, we will summarize and reinterpret the previous experimental results and theories of conventional PCM for a better understanding of X-point PCM. We will focus more on the practical issues and solutions for SCM, rather than having deep theoretical discussions for each part. In each section and part, a change of technical direction will be suggested to satisfy the requirements of SCM. In Section III, more content of this article will be spared for the introduction of X-point PCM.

In Section II, starting with a quick review of the history of PCM, there will be a technical discussion for each of its components: the history of the PCM part, the memory element part, the access device part, and the sensing scheme part. In Section II-A, the different device generations throughout the evolution of PCM will be briefly introduced, mainly for the technode, the integrated structure, and the access device. In Section II-B, there will be a thorough analysis/explanation on the Reset–Set write and thermal disturbance (TDB). In Section II-C, the evolution of the access device from the transistor and diode to the selector for X-point memory will be discussed in the context of the architectural change of PCM. In Section II-D, the sensing schemes of PCM will be explained regarding the latency, drift, and disturbance of PCM, including X-point PCM. In Section III, the basic key technologies on X-point PCM will be described for the first time. The basic operation theory, requirements for the phase change material (PM) and selector, and the issues of process integration will be discussed. The key issues regarding the reliability of X-point characteristics are also included. As a future direction for PCM, vertical X-point (VXP) PCM is also proposed. In Section IV, we will briefly touch on the possibility of PCM for neuromorphic application. The feasibility will be compared with the previous devices in terms of basic cell characteristics.

II. Technical Review on Each Component of PCM

A. Quick Review on the History of PCM

The first demonstration of PCM is as old as that of the silicon transistors [14]. However, due to the complicated physics of PM in its amorphous state [15] and its huge power consumption required for melting and crystallization, the development of PCM for memory application has been static for a long time, despite the fact that chalcogenide materials have found their way in optical storage applications [10]. It was not until the year of 2002 that the revival of PCM was announced by Ovonics-Intel, who demonstrated a 4-Mb array operation with a 180-nm technology node [16].

The cell structure of the first demonstrated PCM was a line-type (L-type) PCM with a high-resistance (R) heater for effective Joule heating, and bipolar junction transistor (BJT) is used as an access device [16]. The cell pitch was 180 nm, and the array density was 4 Mb. After many experimental tests, cell sizes were eventually scaled down to 20 nm [5]. In addition, a variety of access devices, such as the FinFET [17] and the vertical diode [4], were tested in order to provide a better ON-state current (I_{ON}) and OFF-state leakage current (I_{OFF}) ratio (I_{ON}/I_{OFF}) and to reduce the area of the core circuit. In order to increase the bit density of PCM, the concept of the multilevel cell (MLC) was introduced and demonstrated for the first time [18]. Finally, the first mass productions of PCM were successfully made by Samsung [3] and Numonyx (now Micron) [2] with a chip density of 512 Mb and 1 Gb and a cell size of 60–65 and 45 nm, respectively. In both productions, the PMs were patterned as a line shape along the bitline (BL) direction and the bottom electrode functioned as a heater in a ring and wall shape. The shapes of the heater help to decrease I_{Reset} through the increase of R. However, the L-type PCM caused an asymmetric TDB due to the different thermal conductivity of PM (along BL), oxide, and the interface [along the wordline (WL)]. In other words, the L-type resulted in a more severe TDB since the accumulated heat was distributed more along the line direction than the other direction. In order to decrease I_{Reset} and TDB, the confined structures (C-type) of PCM, which were enclosed by oxides in four directions, were successfully developed and

scaled down to 7.5 × 17 nm [6]. However, no more mass productions were made for these new C-type PCMs. We will discuss this PM structure further in Section II-B2.

For the access device, the p-n diode and BJT diode were used for mass production to enhance the I_{ON}/I_{OFF} ratio and minimize the size of the core circuits. However, these diodes required additional via contacts are on every 8th or 32nd cell [3], [4], which increased the areal overhead. Above all else, these access devices did not have 3-D stackability; this will be discussed in Section II-C again. As a result, the people's attention was naturally moved to the stackable amorphous selectors, such as the ovonic threshold switch (OTS) and mixed ion–electron conductor (MIEC). After intensive study and research, Intel-Micron announced the successful development and mass production for stackable 3-D X-point memory using an amorphous selector [19]. SK-Hynix also announced the successful development of stackable X-point PCM using a novel two-terminal selector [9]. The details will be discussed in Sections II-C and III-B.

B. Memory Element (PM)

The history of PCM started with the discovery of chalcogenide alloys, which shows a large and clear R gap and reversible transition between amorphous and (poly-) crystalline phases. Chalcogenide alloys include at least one chalcogen from the group VI-A elements but mostly refers to sulfides, selenides, tellurides, and polonides. The lone pair-induced unique characteristics [20] of the alloy show two unique behaviors: the switching effect and the memory effect [21]. The former is a characteristic of OTS and the latter is a characteristic of PM, which will be discussed here. When compared with that of amorphous silicon (a-Si), the melting temperature (T_{Melt}) of the chalcogenide alloys is much lower; T_{Melt} of a-Si is more than 1300 °C [22], but T_{Melt} of chalcogenide alloys are mostly below 700 °C [23]. In addition, a lone pair-induced unique amorphous structure causes the high density of states around the mobility edge and enables hopping conduction [24]. At the high field, this conductive behavior triggers the threshold switching and causes a transition to the metal-like "ON-state," the so-called insulator–metal transition (IMT) [25]. Several theoretical explanations on the switch mechanism were made: the thermal model [26], carrier generation model [27], and field-driven energy gain model [28]. However, the debates are still inconclusive. The phase change between Reset and Set happens in the "ON-state" region in which current pulse shapes can be controlled by a bias.

Reset operation is about melting and quenching to form an amorphous structure. The practical issues of the Reset operation are I_{Reset} for less power consumption and TDB for low raw bit error rate (RBER). Those are the major traditional issues in PCM and have been discussed for a long time [11]. However, below 2× nm technodes with a C-type PCM, greater importance should be put on the Set operation. While a short and high current pulse [$I >$ melting current (I_{Melt})] makes an amorphous structure, the same high current but with a trailing

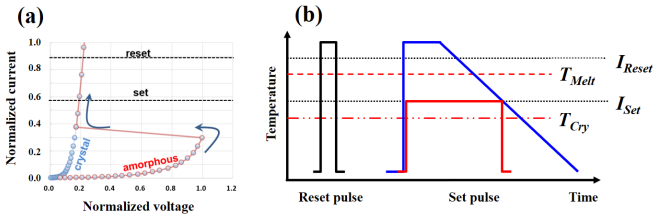

Fig. 2. (a) I–V curve of Set and Reset states based on the field-driven energy gain model [28]. (b) Schematics of the write pulse for Reset and two different Sets, which are MSC and SPC. Note that the MSC consumes more energy than SPC, which is visible from the area under the pulse.

pulse or a long and low current pulse ($I < I_{Melt}$) will lead to crystallization [see Fig. 2(b)].

The first Set method is a crystallization by solidification and will be referred to as "melt and slow cooling" (MSC), and the second Set method is called "solid phase crystallization" (SPC). So far, the difference between MSC and SPC has not been emphasized enough in the past. However, it should be treated more carefully in the SCM application, which requires low write latency for both Set and Reset. In terms of the write latency, because the Set operation takes a much longer time than the Reset operation, there will be an overall delay of the write latency and this will cause the limit of the write bandwidth in the system level.

In this section, the traditional issues of the conventional PCM will be reviewed first, and then, the necessity of a C-type for the SCM will be explained.

1) Reset and Set Write: In this section, we will review the practical issues of Reset and Set write in the conventional L-type PCM with a heater and compare them with the C-type PCM without a heater. Through this comparison, it will be revealed that the development direction of the PM may need to be modified to put more weight on Set write rather than Reset write.

Generally, I_{Reset} is defined as the minimum current needed to complete amorphization. The key issues during the Reset operation were decreasing I_{Reset} and TDB, whereas Reset latency is not problematic because the latency of 20–30 ns is sufficient for any application. However, on the array level, I_{Reset} is associated with many other issues, such as the threshold voltage of the Reset (V_{th_Reset}) distribution, the write endurance, and the TDB. These issues are also the functions of the I_{Reset} distribution.

Traditionally, there have been two ways to reduce I_{Reset}: developing a new chalcogenide alloy and scaling down either the metal heater (for the L-type) or the dimensions of PM (for the C-type). In conventional PM ($Ge_2Sb_2Te_5$), several dopants (N-, SiO_2-, or C-) are known to reduce I_{Reset} [29]–[31]. So far, two different explanations have been made. One is that the doping, or adding dopants, increases the resistivity of PM to increase the efficiency of Joule heating for the same current [29]. The other is that the doping reduces thermal conductivity (k_{th}) of PM, which also makes Joule heating efficient by the thermal confinement effect [30]. Regardless of this mechanism, these dopants cause a serious downside in Set performance, which can be represented by the resistance

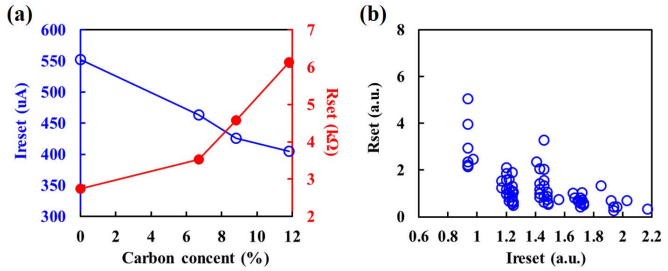

Fig. 3. (a) Tradeoff between I_{Reset} (open circle) and R_{Set} (solid circle) by increasing carbon [31]. (b) Tradeoff between I_{Reset} and Set performance (R_{Set}) in $Ge_2Sb_2Te_5$ [32]; lower R_{Set} represents a faster Set behavior. Tradeoff between I_{Reset} (open circle) and R_{Set} (solid circle) by increasing carbon doping concentration [31].

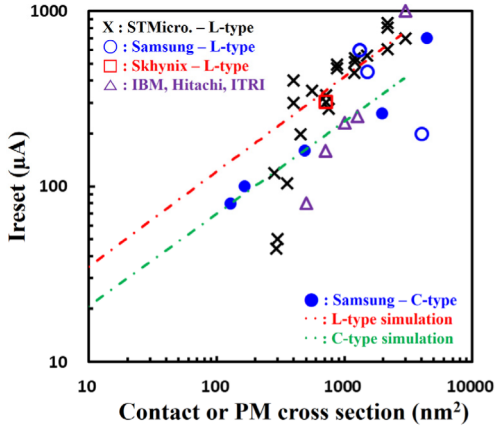

Fig. 4. Scaling trend of I_{Reset} in L-type and C-type PCM. Dotted lines: theoretical models for the two structural types. Each symbol indicates the experimental results from various affiliations [2], [4], [29], [33]–[35].

of Set (R_{Set}). Fig. 3(a) shows the tradeoff by increasing carbon doping [31] and Fig. 3(b) shows the tradeoff between R_{Set} and I_{Reset} by process integration for pure $Ge_2Sb_2Te_5$ [32].

On the other hand, when we look at the overall decreasing scaling trend of I_{Reset} for the C-type in Fig. 4, I_{Reset} is already 50–110 μA at 20-nm technode [6], [33], [34] and it will be even lower in smaller dimensions. If so, the value is already sufficient for SCM application and the efforts to reduce I_{Reset} by adding dopants or developing new alloy systems will not be effective.

Instead, the efforts may have to be made for the Set write. What actually dominates the overall power consumption is not the Reset write but the Set write. In the traditional L-type, the Set write was mostly done by the MSC method, which still needs I_{Reset} for melting. Thus, it was a sufficient way to decrease I_{Reset} for overall power consumption. However, once we adopt the C-type for better scalability and decide to use SPC, the Set write current (I_{Set}) will be smaller than I_{Reset} and the power and/or energy consumption will be dependent on I_{Set} and the pulsewidth (PW). Hence, the development of materials and integration technologies should aim for the reduction of I_{Set} and Set PW.

Comparing SPC with MSC in terms of power consumption and Set characteristics will be another important discussion.

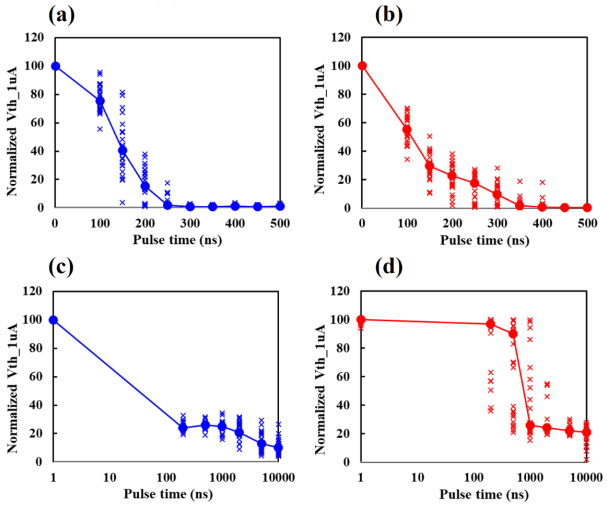

Fig. 5. Normalized $V_{th\_1\ \mu A}$ of $Ge_2Sb_2Te_5$ by Set write pulse time. The effect of SQ and TZ is compared in two different structures. (a) and (b) L-type. (c) and (d) C-type. Here, $V_{th\_1\ \mu A}$ is defined as the voltage at the 1-μA current.

Fig. 2(b) shows the pulse diagram for MSC and SPC. Usually, the crystallization temperature (T_{Cry}) is about 1/2–1/3 of T_{Melt} [36]. This proportionality can also be applied to I_{Set}, which means that SPC consumes less power and energy. For the Set performance, very few reports have made such a direct comparison. Mantegazza *et al.* [37] reported that, in the μ-trench structure, the SPC method by square pulse (SQ) can reduce the Set time (t_{Set}) to 35% from the t_{Set} of trapezoidal pulse (TZ) for the MSC method. We also have the same result in both L- and C-type PCMs. Fig. 5(a)–(d) shows the variations of normalized threshold voltage at 1 μA ($V_{th\_1\ \mu A}$) by the Set write pulse time for $Ge_2Sb_2Te_5$ in both structures; here, $V_{th\_1\ \mu A}$ is defined as the voltage at the 1-μA current. In the L-type, the SQ shows faster Set than the TZ, particularly in the ending stage of the Set, but, in the C-type, the SQ is much faster than the TZ in the overall Set region. The same trend has been found in the array data of the X-point PCM.

Fig. 6(a) compares the threshold voltage of Set (V_{th_Set}) distribution of TZ and SQ for the same 1-μs total PW in the X-point PCM. The SQ shows a much tighter distribution than the TZ and, interestingly, increasing maximum current makes the wider distribution. The result indicates that increasing maximum temperature is not helpful for crystallization, which can be expected in the time-temperature-transformation (TTT) diagram. The possible explanations are that complete melting removes the crystal seed and/or steep cooling leads to more of an amorphous phase.

However, structure-wise t_{Set} of a C-type has no advantage over that of the L-type, though it is necessary for downscaling and TDB. A direct comparison between the two types was made in Fig. 5, where t_{Set} of the C-type is longer than the L-type. A possible explanation is that the C-type has no crystal seeds after full Reset, and its degree of freedom for atomic movement is more limited than that of the L-type [9].

In summary, we discussed the Reset and Set operations of PCM in terms of power and performances. For the

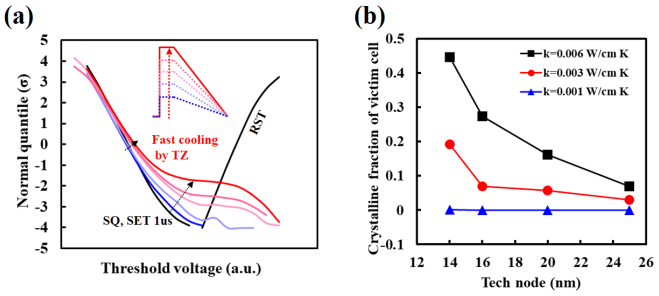

Fig. 6. (a) Distributions of V_{th_Set} and V_{th_Reset} with different current amplitudes in the TZ (blue line to red line) compared with SQ (black line). Note the higher the current amplitude, the wider the V_{th_Set} distribution. (b) Scaling trend of TDB with varying $k_{th,ILD}$. TDB is expressed by the crystalline fraction of the victim cell, which is calculated by electrothermal simulation [43].

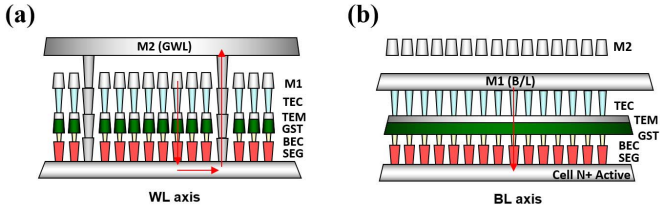

Fig. 7. Schematics of L-type PCM with an SEG diode along (a) WL axis and (b) BL axis. Here, via contacts are on every 8th cell [4].

C-type, which will be the ultimate structure of PCM for SCM application, the burden on I_{Reset} reduction will decrease with downscaling. Instead, the importance of Set performance will increase in terms of power or energy, speed, and distribution. In those three criteria, the SPC method will be more promising than the MSC method.

2) Thermal Disturbance (TDB): It has long been thought that the downscaling of PCM will face a limit from TDB, because the device intrinsically uses high temperature for melting. Negative forecasting has been reported many times [11], [38], but most of the reports are based on L-type structures. In both structures, high temperature from the cell under the Reset operation is transferred along the continuous PM, which has a relatively high k_{th}; in L-type, either the x- or y-direction is separated by an interlayer dielectric (ILD), which has relatively low k_{th}. Redaelli *et al.* [39] reported the experimental result of TDB in a 45-nm L-type PCM and showed asymmetric TDB along the x- and y-directions, not because of the cell-to-cell distance but because of the different structures such as an ILD-bulk and the interface between the PM and the ILD [39]–[41]. Therefore, having these advantages in both the x- and y-directions will be a huge benefit in reducing overall TDB. Lee *et al.* [42] also expected less TDB in a C-type than an L-type.

However, those simulations only showed a temperature gradient from an aggressor to the victim cells. No decent model was proposed to expect the actual variation of R_{Reset} or V_{th_Reset} in the victim cell for the prediction of the ultimate limit of PCM. Recently, Yoo *et al.* [43] reported an electrothermal model combined with a crystallization model for an X-point PCM with a C-type and successfully matched the calculation results with experimental ones. In the report, they predicted TDB for technodes as low as 14 nm with changing $k_{th,ILD}$ [see Fig. 6(b)].

As expected, the TDB increases as technodes get smaller and as $k_{th,ILD}$ increases. The degradation gets especially faster in technodes below 16 nm and comes from the fact that the location of the victim cell is so close to the aggressor that the temperature of the victim cell approaches T_{Melt} faster. This means that, from 14 nm, PCM will face a more serious TDB, and hence, there needs to decrease $k_{th,ILD}$ or make the thermal boundary layer (TBL) [40], [41] or air gap [44] between two PMs. In terms of PM development, adopting high T_{Cry} may be another way to suppress the crystallization of victim cells. However, this will conflict with t_{Set}, which is a necessary condition of SCM. Another method to reduce TDB is to optimize I_{Reset} and reduce the Reset PW as short as possible, which can lower the total heat generation from the aggressor cell.

In summary, in order to scale down the PCM down to 14 nm and below, a variety of efforts should be made. Process-wise, a lower $k_{th,ILD}$, a new TBL, and an air gap will be necessary. For the pulse algorithm, an extreme short Reset PW will be helpful. As with NAND Flash memory, which has an intrinsic problem of charge coupling between the adjacent cells, applying controller technology will be helpful in reducing the RBER caused by TDB between the adjacent cells. It will be another way to reduce TDB and may enable the further downscaling of PCM below 10 nm.

C. Access Device (Selector)

The development history of access devices for PCM can be understood in three ways: capability for high I_{ON}, least I_{OFF}, and better area efficiency for more net die in a wafer. In the very early stages of PCM development, when downscaling was not enough to reduce I_{Reset}, I_{ON} capability was the most important, since large-sized PM required high I_{Reset}. Thus, either a single large-sized MOSFET [17] or dual/triple [45] MOSFETs were adopted. However, these were not suitable for better area efficiency to increase the net die per wafer. Hence, it is natural that a diode using selective epitaxial growth (SEG) came up to replace MOSFETs, because it is capable of high I_{ON} with low I_{OFF} and only takes a small area through the vertical structure.

Because of this breakthrough, industries were able to make the first mass production of PCM [3]. However, the area efficiency was not good enough, since it required via contacts are on every 8th or 32nd cell [3], [4] (see Fig. 7).

In the PCM with a diode structure, a WL signal is transferred to each cell through the doped active region, which has poor conductance but is a necessary outcome for SEG. Therefore, a new solution was proposed to use a metal line, instead of the doped active region [46]. Conceptually, the metal-line-assisted diode should enable a stackable structure, since it does not require single crystal silicon. However, in reality, there is the challenge of the polysilicon diode. Since now the diode must be formed on the metal, the diode will become a polycrystalline structure. Grain boundary-associated leakage and cell-to-cell variations were not easy to control. The idea of a stackable structure seemed impractical, since the diode

TABLE I
PROS AND CONS FOR EACH SELECTOR

	OTS [47]	MIEC [48]	Doped α-Si[49]
Material	AsTeGeSiN	Cu-based	As-SiO_2
I_{off}	10nA@ 0.5 V_{th}	100pA@0.5 V_{th}	20nA@ 0.7 V_{th}
I_{on}	100μA (CC)	10nA (w/o CC)	>100μA
Instability	RTN & *drift*	Small RTN	Huge RTN
V_{th} adjustability	Good	No	Limited
Endurance	$>10^8$	$>10^5$	$>10^5$

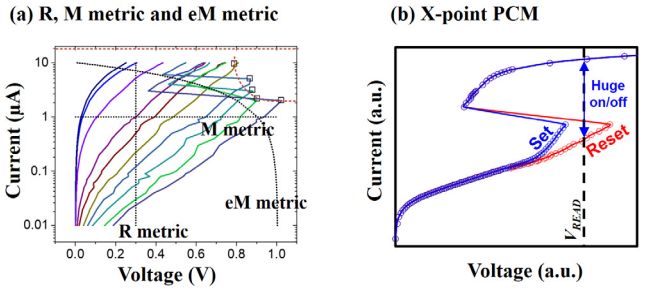

Fig. 8. (a) Three different sensing conditions of R, M, and eM-metrics are shown [55]; constant voltage forcing and current sensing for R-metric, constant current forcing and voltage sensing for M-metric, and varying current forcing near I_{th} and voltage sensing for eM-metric. (b) Suggested constant voltage-forcing and current-sensing scheme for low read latency in X-point PCM.

requires a high-temperature anneal for activation (>700 °C). Such high temperatures will destroy PM on the lower deck, since PM and general chalcogenide alloys are vulnerable in temperatures higher than 350 °C.

In order to further increase the bit density with increasing cost effectiveness, a new stackable access device was required. Amorphous selectors, such as the OTS, the MIEC selector, and the doped α-Si, have been suggested as solutions. The pros and cons of each selector have been summarized in Table I. The research for each selector has been done for a long time, but the efforts were not successful in satisfying the basic requirements of low I_{OFF} and high I_{ON} [50]. The pros and cons of each selector have been summarized in Table I. Recently, industries either announced the readiness of the mass production of memory products using selectors [3] or have even launched the product in the market [19], though the technologies have not been revealed yet. We will review the basic schemes and technologies further in Section III.

D. Sensing Schemes

Since PCM has a large and clear R gap between the Set and Reset states, it is not difficult to judge ON and OFF states of the single-level cell (SLC). The general sensing is done by forcing a constant voltage and sensing the current to compare the resistance, which is named "R-metric" [51]. Usually, a low read voltage ($V_{\mathrm{READ}} \sim 0.3$ V) is applied for a better sensing margin and to avoid read disturbance (RDT). As shown in Fig. 8(a), the lower the read voltage, the larger the current gap between the Set and Reset cells, which will improve the sensing margin. However, a read voltage lower than 0.3 V should be considered for a sensing circuit (in terms of the reference resistance) and a worse drift, which will be a big hurdle for MLC below. On the other hand, increasing V_{READ} will not only reduce the sensing margin but also lead to RDT, due to the unintended turn-on of the low $V_{\mathrm{th\_Reset}}$. This turn-on causes a logic state change from Reset to Set by crystallization, which is an R-metric associated RDT ($\mathrm{RDT_{R-metric}}$) [52].

For MLC, since the total sensing margin is not enough to differentiate the four intermediate logic states of partially amorphized phases [53], the reduction of the drift to minimize the loss of the sensing margin is very important. The drift, which is an increase of V_{th} by defect annihilation over time, is a function of the defect density and an electric field. A reduced drift impact can be expected in high-voltage regions, since the barrier height for electron hopping is lowered at a high field and the impact of the defect density on the drift is reduced [54]. Therefore, it is advantageous to sense the current or voltage near the threshold region ($V_{\mathrm{th\_Reset}}$ and $I_{\mathrm{th\_Reset}}$) for MLC. Considering the wide variation of $V_{\mathrm{th\_Reset}}$, applying 1 μA, which is near $I_{\mathrm{th\_Reset}}$ (3–5 μA), and sensing the voltage will be a safer way to avoid an erratic turn-on of a low $V_{\mathrm{th\_Reset}}$ cell and minimize the drift. This sensing scheme is called "M-metric" for MLC [51] in Fig. 8(a).

However, forcing small currents take a long time to settle the bias at the BL or WL node with large parasitic capacitance, which will increase the latency of MLC sensing. Therefore, an enhanced M-metric named "eM-metric" is proposed to improve the latency by applying maximum current (close to I_{th}) for each cell that has a different V_{th} value and a resulting I_{th} value [55]. These three sensing conditions are shown in Fig. 8(a).

In X-point PCM, as shown in Fig. 8(b), since the combination of PM and selector makes two different I–V curves with different $V_{\mathrm{th\_Set}}$ and $V_{\mathrm{th\_Reset}}$ values, applying a constant read voltage between them will make huge $I_{\mathrm{ON}}/I_{\mathrm{OFF}}$, which will significantly lower the read latency by fast charging of the WL or BL node. However, this will trigger the critical $\mathrm{RDT_{X-point}}$ from the Set state to the Reset state due to the high overshoot current during the snap-back switching. This short and high current is enough to amorphize the Set cell; note that this $\mathrm{RDT_{X-point}}$ is different from $\mathrm{RDT_{R-metric}}$. We will discuss this in Sections III-A and III-E for x-point PCM. Current forcing and voltage sensing, such as the M-metric, may also be possible using the subthreshold region, but it may take an even longer time than the normal M-metric to settle down the WL or BL node, since I_{th} of the selector is generally lower than that of PM.

III. X-POINT PCM FOR SCM

The X-point structure has been thought of as the most ideal structure for memory devices. It has an ultimate cell areal density of $4F^2$ and requires only two photo-mask steps for patterning, which makes it possible to minimize the overall number of process steps. In addition, it is stackable, which enables 3-D structures. Based on these advantages, intensive efforts have been made to achieve this structure. However, only

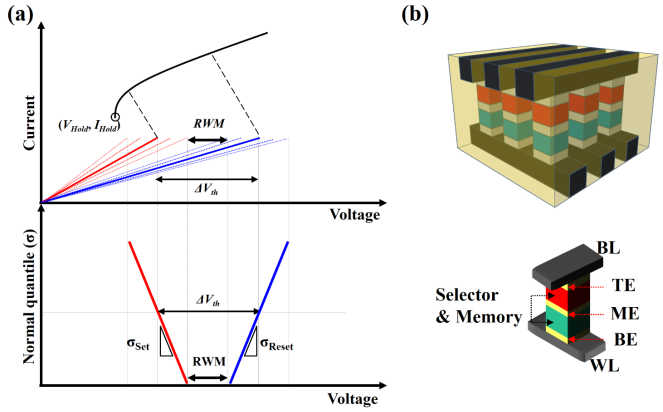

Fig. 9. (a) Variations of I–V curves for Set and Reset and the corresponding V_{th} distributions. The relationship among the V_{th} distributions, ΔV_{th}, and RWM at a low probability region is shown. (b) Schematics for a cell stack of X-point PCM. TE, ME, and BE refer to the top electrode, the middle electrode, and the bottom electrode [9].

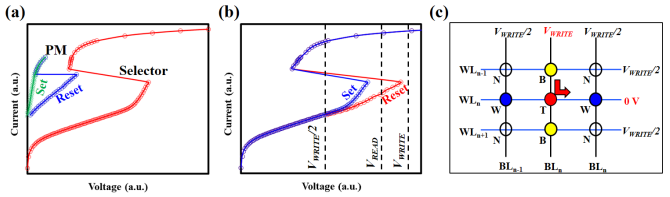

Fig. 10. (a) I–V curves of a selector and PM in a Set and Reset state. (b) I–V curves of X-point PCM when a selector is combined with PM in a Set and Reset state. Read bias (V_{READ}), write bias (V_{WRITE}), and inhibit bias (V_{INH}) are marked. (c) Write bias conditions of the selected and deselected WLs and BLs. Cells are in four different states: T (target cell, V_{WRITE}), B and W (deselected cells in selected BL or WL, $V_{WRITE}/2$), and N (deselected in WL and BL).

a couple of industries [8], [9], [19] seem to be successful in obtaining the fully functional X-point array. In this section, we will review the operations, requirements, and technical hurdles for the development of a fully functional X-point array.

A. Basic Operation Theory

X-point memory array basically consists of a cell stack between two crossing metal lines, called the BL and the WL. Each cell stack includes a memory material and a selector material, and, if necessary, additional electrodes may be inserted to work as a diffusion barrier or to enhance the electrical properties of the two active materials (see Fig. 9).

Since combining a memory element [PM or resistive material (RM)] with a selector in a single stack will make a series connection, the I–V curve will be merged, as shown in Fig. 10(a) and (b).

In the X-point structure that has only two input nodes (WLs and BLs), there is only one way to select and deselect a cell. Fig. 10(c) shows the bias condition to differentiate a target cell ("T") from the other deselected cells in the same WL ("W") and BL ("B"). Generally, the inhibit bias must be half of the read bias ($V_{INH\_READ} = 1/2\ V_{READ}$) or the write bias ($V_{INH\_WRITE} = 1/2\ V_{WRITE}$) to minimize the sneak current (I_{sneak}). Since high-speed sensing requires fast charging, constant voltage forcing and current sensing is

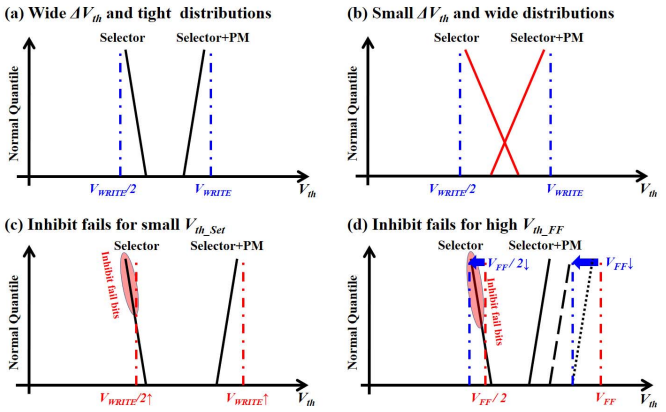

Fig. 11. Schematics for designing various V_{th}(V_{th_Set}, V_{th_Reset}, and V_{th_FF}) and bias conditions (V_{WRITE}, V_{INH}, and V_{FF}). V_{INH} of (a)–(c) is $V_{WRITE}/2$. V_{INH} of (d) is $V_{FF}/2$. (a) Wide ΔV_{th} and tight distributions for a successful bias condition set up. (b) Small ΔV_{th} and wide distributions that lead to no RWM at a low probability region. (c) Inhibit fails for small V_{th_Set}. (d) Inhibit fails due to high V_{th_FF} and the corresponding high V_{INH}.

more advantageous in separating Set cells (after snap-back at V_{READ}) and Reset cells (subthreshold region at V_{READ}) with large I_{ON}/I_{OFF} [see Fig. 10(b)]. Current forcing and voltage sensing are also available in the subthreshold region, but it will take a much longer time to separate the two logic states, which would result in a degradation in read latency.

Normally, V_{th} distributions of amorphous materials, such as PMs, RMs, and selectors, are far worse than those of transistors. Hence, the V_{th} window (ΔV_{th}) of a memory element (PMs or RMs) should be large enough to separate the maximum V_{th_Set} cells (selector V_{th}) and minimum V_{th_Reset} cells (selector V_{th}+ memory V_{th}), as shown in Fig. 11. Thus, when designing the X-point memory, the following sequence is recommended.

1) Estimate the possible V_{th} distributions of Set and Reset.
2) Determine the ΔV_{th}(V_{th_Reset}−V_{th_Set}) to separate the overall Set and Reset states with a sufficient read window margin (RWM).
3) Based on maximum V_{WRITE} and voltage for a first firing (V_{FF}), decide the target V_{th} of the selector to reduce I_{sneak} or to avoid turning on deselected cells by V_{INH_WRITE} or V_{INH_FF}.

Among the procedures, satisfying the conditions of step 3 will be very difficult. Most often, amorphous materials intrinsically require the additional V_{FF} process, which is an initialization for electrical activation. The phenomenon may be considered as a sort of forming process in RRAM. Generally, the reported delta firing voltage ($\Delta V_{th\_FF} = V_{th\_FF}-V_{th\_Set}$) is around 1 V for RRAM [56] and greater than 1 V for OTS. The resulting V_{th_FF} will be additionally increased and the corresponding high V_{INH_FF} ($= 1/2 V_{th\_FF}$) is apt to make high I_{sneak} or even turn on the deselected cells. Hence, combining two materials (RM and OTS) that both have large firing voltages would not be helpful. On the other hand, because as-deposited PM is prone to be a crystalline state, PM has a negligibly small V_{th_FF} that makes it more advantageous. Therefore, in the case of X-point PCM, developing a selector with small ΔV_{th_FF} is very important and the ideal case will

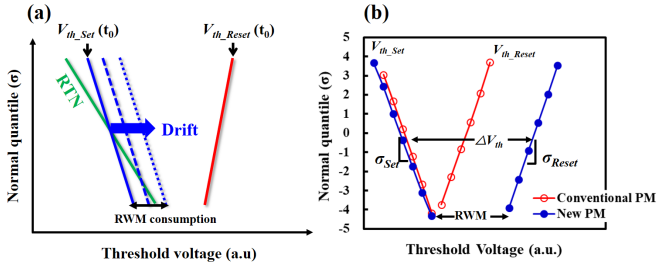

Fig. 12. (a) Impact of RTN and drift on the V_{th} distribution and RWM [9]. (b) Advantage of developing a new PM for a wide ΔV_{th} and RWM [9].

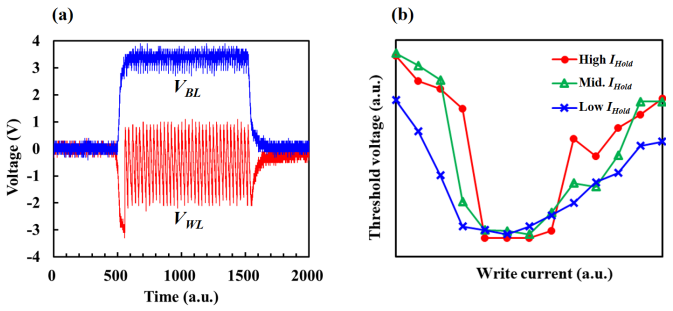

Fig. 13. (a) Oscillation of the voltage in OTS as a function of time when $I_{DUT} < I_{Hold}$. (b) V_{th}–I curve of an X-point PCM with different I_{Hold} values.

be ΔV_{th_FF} that is less than ΔV_{th} of the memory element. Otherwise, another bias condition for V_{th_FF} and V_{INH_FF} should be assigned to the cost of power consumption.

B. Requirements of the Selector

Due to the advantage of X-point memory, there have been intensive efforts to develop a decent selector that is the core of the X-point memory. Table I shows a summary of the various selectors that have been mainly studied so far. The summary shows that each selector has its own "pros and cons," but when applying the necessary practical criteria for a real product, most of the selectors will not be successful. Therefore, in this section, we will discuss the criteria and feasibility of developing a high-density product.

As mentioned in the *access device* part, the first condition for the X-point array is to satisfy both the sufficiently low I_{OFF} and high I_{ON} for read and write operations, because a voltage drop due to I_{sneak} at the crossing point of WL and BL reduces both the voltage and current that reach the far cells. Since PCM uses high I_{Reset}, reducing I_{OFF} of the selector at the given V_{INH} is especially important. A high I_{sneak} value will result in three different issues relative to the degree of insufficiency (see list below). Particularly, the third one may also consume the RWM in the array, even after obtaining the functional die.

1) sensing error that misreads the off-cell for the on-cell;
2) write error from deficient current and voltage supply;
3) uneven distribution of cell V_{th} between the near and far cell.

The second condition to be considered is V_{th} adjustability. As discussed in Section III-A, since we are handling poor uniformity materials, the distribution (i.e., uniformity in a single array) of the materials' R or V_{th} will decide the selectors' optimum V_{th} to avoid I_{sneak}. This indicates that most of the selectors that have a nonadjustable V_{th} will not be applicable for memory elements (PMs or RMs) in a megabit (Mb) size array, considering the V_{th} distribution of each memory element.

V_{th} instability is also a very important condition. It is actually associated with the two conditions mentioned earlier in terms of the V_{th} distribution. Instability means the lack of repeatability of the same cell during the same read or write operation. It is one of the two factors that decide the V_{th} distribution, which are cell-to-cell variability and repeatability. While the former may be related to the intrinsic conduction mechanism and the vulnerability during process integration, the latter is associated with random telegraph noise (RTN) during read. The impact of RTN on the V_{th} distribution and RWM is shown in Fig. 12(a).

The last requirement is the "Hold" characteristic, which represents the minimum current (I_{Hold}) and/or bias (V_{Hold}) to maintain the stable "ON state" of the selector. It is particularly important when the operation current and/or bias is small. Once the operation condition is below the hold values, there will be an oscillation of ON-and-OFF, which is due to the charging and discharging phenomenon [57], [58]. Fig. 13(a) shows the oscillation of the voltage in OTS when the current is below I_{Hold}.

In X-point PCM, I_{Hold} of the selector is particularly important, since it will limit the minimum I_{Set} and corresponding I_{Reset}. Fig. 13(b) shows the V_{th}–I curve of a cell stack with different I_{Hold} values. It shows that the beginning point of the crystallization, which is shown as the onset of V_{th} reduction, is limited by I_{Hold}.

C. Requirements of Memory Material

In the X-point memory array, the most important characteristics of the memory material are a sufficient ΔV_{th} and a decent distribution to secure the RWM at low probability regions. The importance of the two characteristics of low RBER is shown in Fig. 12(b) [9].

Since a cell stack's V_{th} is the sum of the selector and memory element, the V_{th} distribution of both the PM and the selector should be as uniform as possible. The overall RWM can be expressed as follows:

$$\mathrm{RWM} = \Delta V_{th} - q_{array} \times (\sigma_{Set} + \sigma_{Reset}) \quad (1)$$

where ΔV_{th} is the median V_{th} gap between Reset and Set, q_{array} is the array size by normal quantile plot for the target RBER, and σ_{Set} and σ_{Reset} are the distribution slopes (standard deviations) of the V_{th_Set} and V_{th_Reset}, respectively.

In order to satisfy both the low RBER and the low write latency (for high-bandwidth SCM), decent V_{th} distribution should be achieved even with the short write pulse, since a practical write latency will be decided on the condition that the RBER is still satisfactory. Fig. 14(a) and (b) shows the degradation of the V_{th_Set} distribution due to the shorter PW. Careful selection of the materials and well-controlled integration can improve the distribution even in the shorter PW.

In X-point PCM, both the ΔV_{th} and V_{th} distributions are dominated by PM's Set performance. As mentioned in

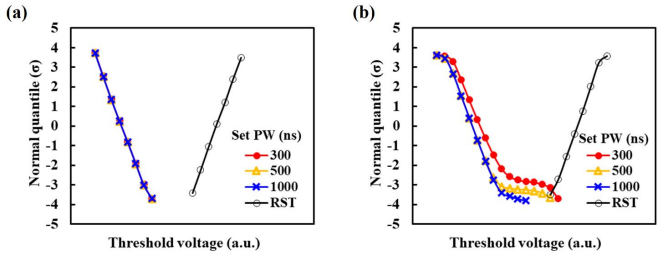

Fig. 14. Change of V_{th_Set} distribution by various SET PW. (a) Well-controlled process integration shows no difference in V_{th_Set} by decreasing SET PW. (b) Unoptimized integration results in a wider V_{th_Set} distribution with growing tail bits by decreasing SET PW.

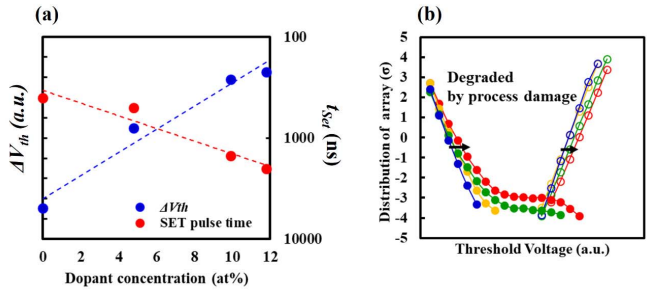

Fig. 15. (a) Tradeoff between ΔV_{th} and t_{Set} of PM by increasing E_G (by increasing dopant concentration) [9]. (b) Degradation of the V_{th_Set} and the V_{th_Reset} distribution by process integration damage. Notice a relatively smaller degradation in the V_{th_Reset} distribution than V_{th_Set} distribution [9].

Section II-B1, Set write takes a much longer time than Reset write. Generally, ΔV_{th} and t_{Set} are in a tradeoff relationship. It is because PM's V_{th_Reset} is a function of bandgap (E_G). The wider E_G, the stronger atomic bonding it has. On the other hand, stronger bonding requires more energy for bond breaking and rearrangement for crystallization. Hence, developing a new alloy system or adding dopants of wide E_G materials is likely to result in longer t_{Set}, as shown in Fig. 15(a). Since faster t_{Set} will also be very helpful for a uniform V_{th_Set} distribution, overcoming the tradeoff between ΔV_{th} and t_{Set} will be the key issue in X-point PCM.

V_{th_Reset} distribution is a direct function of the write current for Reset. While over-Reset will guarantee a uniform V_{th_Reset} distribution, it will cause damage in both the PM and SM, which leads to poor write endurance and TDB. Therefore, to prevent the over-Reset, a uniform I_{Reset} distribution will be helpful, since it will minimize the over-Reset of the cells with small I_{Reset}, when the constant I_{Reset} is applied. For the same reason, in the L-type PCM, where the distribution of I_{Reset} was not uniform, the write-and-verify (WnV) scheme had to be used to minimize the over-Reset [59], which will improve TDB and endurance, at the cost of high latency and power consumption. In contrast, because X-point PCM has a C-type, it is more advantageous in terms of the I_{Reset} distribution. As shown in Fig. 12(b), a uniform V_{th_Reset} distribution with a single pulse is sufficient to go without WnV.

D. Process Integration for X-Point PCM

Theoretically, an X-point PCM array can simply be integrated by using the self-align etch scheme [9], since it only requires two photo-mask steps and two self-align etch steps. However, in reality, the handling of chalcogenide alloys is extremely difficult. The alloys genetically include about 20% of structural defects [60], which are formed by lone pairs. Therefore, physical and electrical properties of the rather porous structures were expected to be easily damaged when the standard etch and cleaning processes were applied. In order to avoid this issue, intensive research has been done for the patterning of PM by a "damascene" process using chemical vapor deposition (CVD) [42], atomic layer deposition (ALD) [61]–[64], and physical vapor deposition (PVD) with thermal reflow [65]. However, in addition to their compositions being uncontrollable, those films are even more porous. Especially in the downscaled damascene process, which has a high aspect ratio deep trench, the porousness of the films is actually visible [61]. Other than that, the increasing number of process steps and slow throughput will be big challenges for mass production. Hence, taking the developmental pathway either toward the self-align etch scheme or the damascene scheme will depend on the maturity of each of the technology as well as the electrical properties and low process cost.

Fig. 15(b) shows how the "self-align etch" process integration degrades the distribution of V_{th_Set} and V_{th_Reset}. While V_{th_Reset} distributions show a slight slope change, V_{th_Set} distributions become wider and even form serious tail distributions [9], though no void, defect, or composition change was shown in the image by a transmission electronic microscope (TEM). The results show that Set characteristics are much more vulnerable to the process damages and that the reconfiguration of the atomic bonding for the low R is very sophisticated.

According to our experimental attempts, a similar situation happened in RRAM with the transition metal oxide (TMO). Since the regular process integrations include many oxygen environments such as etching, cleaning, and ozone plasma, the stoichiometry and resulting electrical properties of the TMO-type RRAM ended up with a wide range of R distributions when the device was exposed to the self-align etch scheme.

E. Reliability and Disturbance in X-Point PCM

In terms of the memory array characteristics, the most conspicuous difference of X-point PCM from the traditional L-type PCM is the overshoot current, which happens during the abrupt turn-on of the selector. This overshoot current can only be reduced rather than completely removed, since it originates from the fast swiping charges that are accumulated in the overall capacitance. This abrupt and fast-flowing charge could be an advantage in terms of the fast sensing for read and effective current delivery for write. However, this will cause many side effects for array operations and reliabilities.

The amount of overshoot current is a function of the R and capacitance. Particularly in a large array, cells that are near or far from the write driver will experience different overshoot currents due to R differences. The relationship between the I–V curve and the corresponding I–time curve is shown in Fig. 16(a) and (b). Different R values result in different dynamic slopes in the I–V curve [shown in

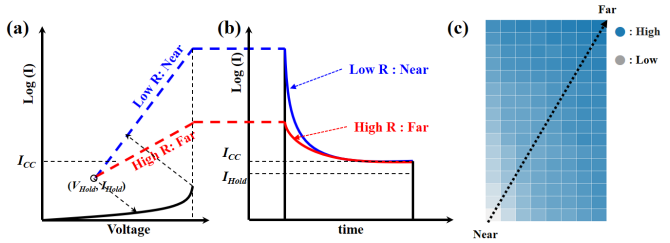

Fig. 16. Relationship between the *I–V* curve and the corresponding *I*–time curve. *(a)* Mechanism of overcurrent generation in the *I–V* curve for two different loadline resistances. (b) *I*–time curve. *(c)* Gradient of V_{th_Set} in a large array of X-point PCM, which is due to the overcurrent difference. Note that the different overcurrents are due to the different loadline resistance at the near and far cell as shown in (a).

Fig. 16(a)], which will cause different overshoot currents in the time domain [shown in Fig. 16(b)]. This signifies that even in a single array, different amounts of overcurrent stress will be forced in each region of the cell: near, middle, and far from the via contact. This difference of stress will result in different Set characteristics and reliabilities, such as cyclic write endurance, $RDT_{X-point}$, and TDB.

Fig. 16(c) shows the gradient of V_{th_Set} in a large array of X-point PCM, which is due to the different stress resulting from the first firing in the array. The larger overcurrent happens at nearer cells during the first firing process and lower V_{th_Set} also forms, but this is beyond the scope of this article.

In order to minimize the overcurrent, a high R and small capacitance will be helpful. However, in PCM, since high R may end up with an insufficient I_{Reset} delivery to the far cells (which will cause Reset failure or a wide V_{th_Reset} distribution), optimum R of WL and BL needs to be calculated. Decreasing overall capacitance in the cell array and circuits will also be helpful. Therefore, it is better to choose the metal with the lowest sheet resistance for both WL and BL so that the metal can provide sufficient current for Reset operation with the minimum capacitance between two metal lines. A similar but even worse problem may happen in X-point RRAM, especially with the TMO type. This is because RRAM is reported to be so vulnerable to the current stress that it results in the unstable Set and Reset states and even in soft or hard failures [66].

IV. Future Outlook and Perspective on PCM

So far, we have made technical discussions on the PCM from the conventional L-type through the C-type to the X-point structure. In this section, we will discuss the new 3-D structure of X-point PCM, VXP PCM for higher density SCM, and new rising application as a synaptic device for a neuromorphic calculation.

A. VXP Memory

Although X-point PCM has an ideal areal density of $4F^2$, additional stacking of X-point PCM will increase the total number of photo steps and process steps, which will decrease the cost effectiveness of the memory device and lowers the throughput for mass production. Therefore, a new structure resembling 3-D NAND will be the future direction of the X-point PCM. However, this structure will require ALD processes for PM, selector, and electrode that can separate the two materials physically. The research on ALD-PM has been done for a long time [61]–[64]; however, there are not many reports on ALD-selector. Recently, there was a report about ALD-OTS to make the VXP PCM, as shown in Fig. 17 [67]. The basic structure is the same as in 3-D NAND [68], except that the cell stack for charge trapping is replaced by PM and OTS, as shown in Fig. 17.

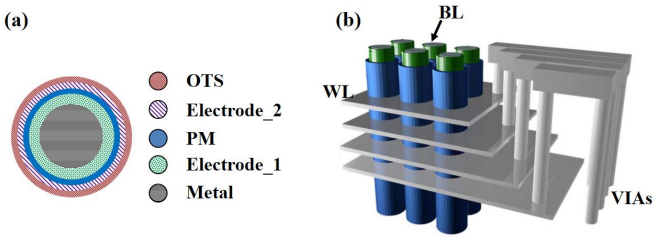

Fig. 17. First proposed VXP architecture. (a) Expected cross section of the BL pillar that shows the whole cell stack. (b) Overall 3-D architecture of VXP PCM [67].

TABLE II
Statistical Consistency and Nonideality of NVMs for a Synaptic Device

	PCM [69], [71]	RRAM [72]-[75]	STT-MRAM [76], [77]	3D-NAND [78], [79]
RTN	Small	Large	Small	Large
Instability (Drift)	Yes	No	No	Yes
Cell to cell uniformity	Good	Bad	Good	Bad
On/off ratio for multi-bit	~ 100s	~ 20-30	< 2.5	> 1E6
Linearity	Good	Some	Some	Good
Symmetricity	N/A	Some	Some	N/A

In this architecture, however, the areal bit density will be low since the cell stack height in X-point PCM will become the area for VXP, as shown in Fig. 17(a). Consequently, the total number of WL stacks may have to be increased to satisfy the total 3-D bit density. Decreasing the cell stack height, which is the sum of the thickness of PM, SM, and electrodes, maybe a solution but will be a difficult task. Since the PM and the OTS are voltage-dependent devices, shrinkage of the thickness will change many device parameters. Therefore, in order to make a decent VXP PCM, it is recommended to develop new materials to shrink the cell stack height or to propose a new 3-D architecture to improve the areal bit density.

B. PCM for Neuromorphic Application

The most interesting future application of PCM is probably the synaptic device for neuromorphic computing [69]–[71]. Intensive studies are being carried out for the synaptic device by using various NVMs, such as PCM [69], [71], RRAM [72]–[75], STT-MRAM [76], [77], and Flash memory [78], [79] (Table II). Although each memory has its own "pros and cons," the kernels of the conditions for synaptic function are the "gradual conductance variation" to emulate plasticity of human brain and statistical consistency in a

memory array, which means the repeatability of read and write, small "cell-to-cell variation" in an array, and large memory window for analog-like multistates; general recommendation is 256 states for 8 bits [80], [81]. Even though neuromorphic computing provides some variation tolerance, the statistical consistency should be still considered seriously; once the consistency is satisfied, other NVM characteristics may be considered as assistant parameters for better performances. Therefore, it should be emphasized that even though there are many reports that show some plasticity [82]–[84], their statistical consistency should be examined, if aiming for the real product. Furthermore, most of the reports show poor linearity and symmetry as well as a small memory window.

In these points of view, PCM is the only device of which statistical consistency is verified with a large memory window in a large array; actually, this is the point why PCM was successful to be a real product. For the synaptic characteristics, a gradual conductance increase can be made by applying the Set pulses of the same amplitude. Although the gradual conductance reduction still needs the incremental step pulse programming (ISPP) for Reset write, it may be avoided by a two-PCM synapse scheme, which may also provide a better symmetry in conductance variation [69], [72]. However, the intrinsic weakness of PCM, which is drift, remains as a hurdle, though there exist some algorithms to minimize this [69], [85], [86].

V. Conclusion

In this article, we introduced the development history of PCM and reinterpreted the technical hurdles for the SCM application. In order to satisfy the SCM requirements, cost effectiveness, power consumption, and performances should be improved and well balanced. The development of X-point PCM was suggested as a new breakthrough to overcome the technical limits of conventional PCM and to satisfy the SCM's requirements. Basic operation theory, integration hurdles, and disturbance of X-point PCM were introduced for the first time. VXP was suggested for the future direction of X-point PCM for better cost effectiveness. However, developing the ALD process and decreasing stack height were expected as the remaining technical hurdles. In the end, a synaptic device for neuromorphic computing was proposed as a promising future application of PCM. The gradual conductance variation and the great statistical consistency of PCM may be a realistic solution for the synaptic device.

References

[1] *International Technology Roadmap for Semiconductors 2007 Edition, Emerging Research Devices*, Int. Technol. Roadmap Semicond., London, U.K., 2007.

[2] G. Servalli, "A 45 nm generation phase change memory technology," in *IEDM Tech. Dig.*, Dec. 2009, pp. 103–106.

[3] P. Clarke. (Dec. 2, 2010). *Phase-Change Memory Found in Handset*. [Online]. Available: https://www.eetimes.com/document.asp?doc_id=1258042#

[4] J. Oh *et al.*, "Full integration of highly manufacturable 512 Mb PRAM based on 90 nm technology," in *IEDM Tech. Dig.*, Dec. 2006, pp. 515–518, doi: 10.1109/IEDM.2006.346905.

[5] Y. Choi *et al.*, "A 20 nm 1.8 V 8 Gb PRAM with 40 MB/s program bandwidth," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Paper*, Feb. 2012, pp. 46–48, doi: 10.1109/ISSCC.2012.6176872.

[6] I. Kim *et al.*, "High performance PRAM cell scalable to sub-20 nm technology with below $4F^2$ cell size, extendable to DRAM applications," in *Proc. Symp. VLSI Technol.*, Jun. 2010, p. 203, doi: 10.1109/VLSIT.2010.5556228.

[7] S.-H. Lee, "Technology scaling challenges and opportunities of memory devices," in *IEDM Tech. Dig.*, Dec. 2016, doi: 10.1109/IEDM.2016.7838026.

[8] D. Kau *et al.*, "A stackable cross point phase change memory," in *IEDM Tech. Dig.*, Dec. 2009, pp. 1–4, doi: 10.1109/IEDM.2009.5424263.

[9] T. Kim *et al.*, "High-performance, cost-effective 2z nm two-deck cross-point memory integrated by self-align scheme for 128 Gb SCM," in *IEDM Tech. Dig.*, Dec. 2018, pp. 37.1.1–37.1.4, doi: 10.1109/IEDM.2018.8614680.

[10] G. W. Burr *et al.*, "Recent progress in phase-change memory technology," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 6, no. 2, pp. 146–162, Jun. 2016, doi: 10.1109/JETCAS.2016.2547718.

[11] S. J. Ahn *et al.*, "Reliability perspectives for high density PRAM manufacturing," in *IEDM Tech. Dig.*, Dec. 2011, pp. 12.6.1–12.6.4, doi: 10.1109/IEDM.2011.6131542.

[12] G. Navarro *et al.*, "Phase-change memory: Performance, roles and challenges," in *Proc. Int. Memory Workshop (IMW)*, 2018, doi: 10.1109/IMW.2018.8388845.

[13] N. A. Goryunova and B. T. Kolomiets, "Electrical properties and structure in system of selenide of Tl, Sb, and As," *Zhurn. Techn. Fiz*, vol. 28, p. 984, 1955.

[14] R. G. Neale, D. L. Nelson, and G. E. Moore, "Nonvolatile and reprogrammable, the read-mostly memory is here," *Electronics*, vol. 43, no. 20, pp. 56–60, Sep. 1970.

[15] D. Adler, M. S. Shur, M. Silver, and S. R. Ovshinsky, "Threshold switching in chalcogenide-glass thin films," *J. Appl. Phys.*, vol. 51, pp. 3289–3309, Jun. 1980, doi: 10.1063/1.328036.

[16] M. Gill, T. Lowrey, and J. Park, "Ovonic unified memory—A high-performance nonvolatile memory technology for stand-alone memory and embedded applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jun. 2003, pp. 202–203, doi: 10.1109/ISSCC.2002.993006.

[17] W. Y. Cho *et al.*, "A 0.18-μm 3.0-V 64-Mb nonvolatile phase-transition random access memory (PRAM)," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 293–300, Jan. 2005, doi: 10.1109/JSSC.2004.837974.

[18] S. R. Ovshinsky, P. K. Klersy, and D. A. Strand, "Electrically erasable directly overwritable multibit single cell memory elements and arrays fabricated therefrom," U.S. Patent 005 296 716 A, Mar. 22, 1994.

[19] P. Clarke. (Jul. 28, 2015). *Intel, Micron Launch 'Bulk-Switching' ReRAM*. [Online]. Available: https://www.eetimes.com/document.asp?doc_id=1327289

[20] M. Kastner, D. Adler, and H. Fritzsche, "Valence-alternation model for localized gap states in lone-pair semiconductors," *Phys. Rev. Lett.*, vol. 37, no. 22, pp. 1504–1507, Jul. 2002, doi: 10.1103/physrevlett.37.1504.

[21] N. A. Bogoslovskiy and K. D. Tsendin, "Physics of switching and memory effects in chalcogenide glassy semiconductors," *Semiconductors*, vol. 46, no. 5, pp. 559–590, May 2012, doi: 10.1134/s1063782612050065.

[22] S. A. Kokorowski, G. L. Olson, J. A. Roth, and L. D. Hess, "Investigation of the melting temperature of amorphous silicon," *Phys. Rev. Lett.*, vol. 48, no. 7, pp. 498–501, Jul. 2002, doi: 10.1103/physrevlett.48.498.

[23] N. Yamada, E. Ohno, K. Nishiuchi, and N. Akahira, "Rapid-phase transitions of GeTe-Sb_2Te_3 pseudobinary amorphous thin films for an optical disk memory," *J. Appl. Phys.*, vol. 69, no. 5, pp. 2849–2856, 1991, doi: 10.1063/1.348620.

[24] N. F. Mott and E. A. Davis, *Electronic Processes in Non-Crystalline Materials*. Oxford, U.K.: Clarendon, 1967.

[25] M. Imada, A. Fujimori, and Y. Tokura, "Metal-insulator transitions," *Rev. Mod. Phys.*, vol. 70, no. 4, pp. 1039–1262, 1998, doi: 10.1103/RevModPhys.70.1039.

[26] D. L. Eaton, "Electrical conduction anomaly of semiconducting glasses in the system As-Te-I," *J. Amer. Ceram. Soc.*, vol. 47, no. 11, pp. 554–558, Nov. 1964, doi: 10.1111/j.1151-2916.1964.tb13816.x.

[27] A. Pirovano, A. Lacaita, A. Benvenuti, F. Pellizzer, and R. Bez, "Electronic switching in phase-change memories," *IEEE Trans. Electron Devices*, vol. 51, no. 3, pp. 452–459, Mar. 2004, doi: 10.1109/TED.2003.823243.

[28] D. Ielmini, "Threshold switching mechanism by high-field energy gain in the hopping transport of chalcogenide glasses," *Phys. Rev. B, Condens. Matter*, vol. 78, no. 3, 2008, Art. no. 035308, doi: 10.1103/PhysRevB.78.035308.

[29] Y. Hwang *et al.*, "Writing current reduction for high-density phase-change RAM," in *IEDM Tech. Dig.*, Mar. 2004, p. 893, doi: 10.1109/IEDM.2003.1269422.

[30] S. W. Ryu *et al.*, "SiO_2 doped $Ge_2Sb_2Te_5$ thin films with high thermal efficiency for applications in phase change random access memory," *Nanotechnology*, vol. 22, May 2011, Art. no. 254005, doi: 10.1088/0957-4484/22/25/254005.

[31] J. H. Park *et al.*, "Reduction of RESET current in phase change memory devices by carbon doping in GeSbTe films," *J. Appl. Phys.*, vol. 117, no. 11, Mar. 2015, Art. no. 115703, doi: 10.1063/1.4914909.

[32] D. Ha and K. Kim, "Recent advances in high density phase change memory (PRAM)," in *Proc. Int. Symp. VLSI Technol., Syst. Appl. (VLSI-TSA)*, 2007, pp. 1–4.

[33] D. Im *et al.*, "A unified 7.5 nm dash-type confined cell for high performance PRAM device," in *IEDM Tech. Dig.*, Dec. 2008, pp. 1–4, doi: 10.1109/IEDM.2008.4796654.

[34] M. J. Kang *et al.*, "PRAM cell technology and characterization in 20 nm node size," in *IEDM Tech. Dig.*, Dec. 2011, pp. 3.1.1–3.1.4, doi: 10.1109/IEDM.2011.6131478.

[35] H.-S. P. Wong *et al.*, "Phase change memory," *Proc. IEEE*, vol. 98, no. 12, pp. 2201–2227, Dec. 2011., doi: 10.1109/JPROC.2010.2070050.

[36] J. Kalb, F. Spaepen, and M. Wuttig, "Calorimetric measurements of phase transformations in thin films of amorphous Te alloys used for optical data storage," *J. Appl. Phys.*, vol. 93, no. 5, pp. 2389–2393, Mar. 2003, doi: 10.1063/1.1540227.

[37] D. Mantegazza, D. Ielmini, E. Varesi, A. Pirovano, and A. L. Lacaita, "Statistical analysis and modeling of programming and retention in PCM arrays," in *IEDM Tech. Dig.*, Dec. 2007, pp. 311–314, doi: 10.1109/IEDM.2007.4418933.

[38] S. Lee *et al.*, "Programming disturbance and cell scaling in phase change memory: For up to 16 nm based $4F^2$ cell," in *Proc. Symp. VLSI Technol.*, Jun. 2010, pp. 199–200, doi: 10.1109/VLSIT.2010.5556226.

[39] A. Redaelli *et al.*, "Interface engineering for thermal disturb immune phase change memory technology," in *IEDM Tech. Dig.*, Dec. 2013, pp. 30.4.1–30.4.4, doi: 10.1109/IEDM.2013.6724724.

[40] J. Reifenberg, D. Kencke, and K. Goodson, "The impact of thermal boundary resistance in phase-change memory devices," *IEEE Electron Device Lett.*, vol. 29, no. 10, pp. 1112–1114, Oct. 2008, doi: 10.1109/LED.2008.2003012.

[41] S. W. Fong, C. M. Neumann, and H.-S.-P. Wong, "Dual-layer dielectric stack for thermally-isolated low-power phase-change memory," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2017, pp. 1–4, doi: 10.1109/IMW.2017.7939079.

[42] J. Lee *et al.*, "Highly scalable phase change memory with CVD GeSbTe for sub 50 nm generation," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2007, pp. 102–103, doi: 10.1109/VLSIT.2007.4339744.

[43] S. Yoo, H. D. Lee, S. Lee, H. Choi, and T. Kim, "Electro-thermal model for thermal disturbance in cross-point phase change memory," *IEEE Trans. Electron Devices*, to be published.

[44] K. S. Choi, "Electronic device and method for fabricating the same," U.S. Patent 2019/0 181 341 A1, Jun. 19, 2019.

[45] S. Kang *et al.*, "A 0.1-μm 1.8-V 256-Mb phase-change random access memory (PRAM) with 66-MHz synchronous burst-read operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 210–218, Jan. 2007, doi: 10.1109/JSSC.2006.888349.

[46] S. Lee *et al.*, "Highly productive PCRAM technology platform and full chip operation: Based on $4F^2$ (84 nm pitch) cell scheme for 1 Gb and beyond," in *IEDM Tech. Dig.*, Dec. 2011, pp. 47–50, doi: 10.1109/IEDM.2011.6131480.

[47] M.-J. Lee *et al.*, "Highly-scalable threshold switching select device based on chaclogenide glasses for 3D nanoscaled memory arrays," in *IEDM Tech. Dig.*, Dec. 2012, pp. 2.6.1–2.6.4, doi: 10.1109/IEDM.2012.6478966.

[48] G. W. Burr *et al.*, "Large-scale (512 k bit) integration of multilayer-ready access-devices based on mixed-ionic-electronic-conduction (MIEC) at 100% yield," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2012, pp. T41–T42, doi: 10.1109/VLSIT.2012.6242451.

[49] S. G. Kim *et al.*, "Breakthrough of selector technology for cross-point 25-nm ReRAM," in *IEDM Tech. Dig.*, Dec. 2017, pp. 2.1.1–2.1.4, doi: 10.1109/IEDM.2017.8268309.

[50] G. W. Burr *et al.*, "Access devices for 3D crosspoint memory," *J. Vac. Sci. Technol. B, Nanotechnol. Microelectron., Mater., Process., Meas., Phenomena*, vol. 32, no. 4, Jul. 2014, Art. no. 040802, doi: 10.1116/1.4889999.

[51] A. Sebastian, N. Papandreou, A. Pantazi, H. Pozidis, and E. Eleftheriou, "Non-resistance-based cell-state metric for phase-change memory," *J. Appl. Phys.*, vol. 110, no. 8, Oct. 2011, Art. no. 084505, doi: 10.1063/1.3653279.

[52] N. Ciocchini and D. Ielmini, "Pulse-induced crystallization in phase-change memories under set and disturb conditions," *IEEE Trans. Electron Devices*, vol. 62, no. 3, pp. 847–854, Mar. 2015, doi: 10.1109/TED.2015.2389895.

[53] T. Nirschl *et al.*, "Write strategies for 2 and 4-bit multi-level phase-change memory," in *IEDM Tech. Dig.*, Dec. 2007, pp. 461–464, doi: 10.1109/IEDM.2007.4418973.

[54] D. Ielmini, D. Sharma, S. Lavizzari, and A. L. Lacaita, "Physical mechanism and temperature acceleration of relaxation effects in phase-change memory cells," in *Proc. IEEE Int. Rel. Phys. Symp.*, Apr. 2008, pp. 597–603, doi: 10.1109/RELPHY.2008.4558952.

[55] J. Cheon *et al.*, "Non-resistance metric based read scheme for multi-level PCRAM in 25 nm technology," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2015, pp. 1–4, doi: 10.1109/CICC.2015.7338358.

[56] H. Akinaga and H. Shima, "Resistive random access memory (ReRAM) based on metal oxides," *Proc. IEEE*, vol. 98, no. 12, pp. 2237–2251, Dec. 2010, doi: 10.1109/JPROC.2010.2070830.

[57] P. E. Schmidt and R. C. Callarotti, "Theoretical and experimental study of the operation of ovonic switches in the relaxation oscillation mode. I. The charging characteristic during the off state," *J. Appl. Phys.*, vol. 55, no. 8, pp. 3144–3147, Apr. 1984, doi: 10.1063/1.333314.

[58] M. Nardone, V. G. Karpov, and I. V. Karpov, "Relaxation oscillations in chalcogenide phase change memory," *J. Appl. Phys.*, vol. 107, no. 5, Mar. 2010, Art. no. 054519, doi: 10.1063/1.3329387.

[59] K.-J. Lee *et al.*, "A 90 nm 1.8 V 512 Mb diode-switch PRAM with 266 MB/s read throughput," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 43, Feb. 2007, pp. 150–162, doi: 10.1109/ISSCC.2007.373499.

[60] T. Nonaka, G. Ohbayashi, Y. Toriumi, Y. Mori, and H. Hashimoto, "Crystal structure of GeTe and $Ge_2Sb_2Te_5$ meta-stable phase," *Thin Solid Films*, vol. 370, pp. 258–261, Jul. 2000, doi: 10.1016/S0040-6090(99)01090-1.

[61] M. B. Sky *et al.*, "Crystalline-as-deposited ALD phase change material confined PCM cell for high density storage class memory," in *IEDM Tech. Dig.*, Dec. 2015, pp. 3.6.1–3.6.4, doi: 10.1109/IEDM.2015.7409621.

[62] J. Lee, S. Choi, C. Lee, Y. Kang, and D. Kim, "GeSbTe deposition for the PRAM application," *Appl. Surf. Sci.*, vol. 253, no. 8, pp. 3969–3976, Feb. 2007, doi: 10.1016/j.apsusc.2006.08.044.

[63] D. Kil, S. Kang, K. Hong, and H. Kim, "The investigation on the gap fill of ALD GST film and new phase change material for 20 nm level PCRAM development," in *Proc. Eur. Phase Change Ovonix Sci. Symp.*, 2013. [Online]. Available: https://www.epcos.org

[64] H. K. Kim, S. J. Chae, B. K. Lee, and H. Kim, "Development of high density PCRAM cell using ALD GST with novel precursors," in *Proc. Eur. Phase Change Ovonix Sci. Symp.*, 2015. [Online]. Available: https://www.epcos.org

[65] M. J. Kang, Y. H. Ha, D. H. Park, J. H. Park, and H. J. Shin, "Phase change memory device and method of fabricating the same," U.S. Patent 2009/0 035 514 A1, Feb. 5, 2009.

[66] Y.-T. Chung, P.-C. Su, W.-J. Lin, M.-C. Chen, and T. Wang, "SET/RESET cycling-induced trap creation and SET-disturb failure time degradation in a resistive-switching memory," *IEEE Trans. Electron Devices*, vol. 63, no. 6, pp. 2367–2373, Jun. 2016, doi: 10.1109/TED.2016.2555333.

[67] V. K. Narasimhan, V. Adinolfi, L. Cheng, and M. E. McBriarty, "Physical and electrical characterization of ALD chalcogenide materials for 3D memory applications," presented at the Int. Conf. At. Layer Deposition (ALD), Jul. 22, 2019.

[68] K. J. Yoon, Y. Kim, and C. S. Hwang, "What will come after V-NAND—Vertical resistive switching memory?" *Adv. Electron. Mater.*, 2019, Art. no. 1800914, doi: 10.1002/aelm.201800914.

[69] S. Ambrogio *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, pp. 60–67, Jun. 2018, doi: 10.1038/s41586-018-0180-5.

[70] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Nov. 2015, doi: 10.1109/TED.2015.2439635.

[71] I. Boybat *et al.*, "Neuromorphic computing with multi-memristive synapses," *Nature Commun.*, vol. 9, no. 1, p. 2514, 2018, doi: 10.1038/s41467-018-04933-y.

[72] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, pp. 61–64, May 2015, doi: 10.1038/nature14441.

[73] P. Yao *et al.*, "Face classification using electronic synapses," *Nature Commun.*, vol. 8, p. 15199, May 2017, doi: 10.1038/ncomms15199.

[74] C. Li *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature Electron.*, vol. 1, no. 1, pp. 52–59, Jan. 2018, doi: 10.1038/s41928-017-0002-z.

[75] F. Cai *et al.*, "A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations," *Nature Electron.*, vol. 2, no. 7, pp. 290–299, Jul. 2019, doi: 10.1038/s41928-019-0270-x.

[76] A. F. Vincent *et al.*, "Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 2, pp. 166–174, Apr. 2015, doi: 10.1109/TBCAS.2015.2414423.

[77] W. Wu, X. Zhu, S. Kang, K. Yuen, and R. Gilmore, "Probabilistically programmed STT-MRAM," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 1, pp. 42–51, Mar. 2012, doi: 10.1109/JETCAS.2012.2187401.

[78] S. Brink, S. Nease, and P. Hasler, "Computing with networks of spiking neurons on a biophysically motivated floating-gate based neuromorphic integrated circuit," *Neural Netw.*, vol. 45, pp. 39–49, Sep. 2013, doi: 10.1016/j.neunet.2013.02.011.

[79] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 μm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, Jan. 2015, doi: 10.1109/JSSC.2014.2356197.

[80] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6869–6898, Sep. 2016. [Online]. Available: https://arxiv.org/abs/1609.07061

[81] B. Jacob *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, doi: 10.1109/CVPR.2018.00286.

[82] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nature Electron.*, vol. 1, no. 1, pp. 22–29, Jan. 2018, doi: 10.1038/s41928-017-0006-8.

[83] Q. Xia and J. J. Yang, "Memristive crossbar arrays for brain-inspired computing," *Nature Mater.*, vol. 18, no. 4, pp. 309–323, Apr. 2019, doi: 10.1038/s41563-019-0291-x.

[84] D. Ielmini and H.-S.-P. Wong, "In-memory computing with resistive switching devices," *Nature Electron.*, vol. 1, no. 6, pp. 333–343, Jun. 2018, doi: 10.1038/s41928-018-0092-2.

[85] D. Lencer, M. Salinga, and M. Wuttig, "Design rules for phase-change materials in data storage applications," *Adv. Mater.*, vol. 23, no. 18, pp. 2030–2058, May 2011, doi: 10.1002/adma.201004255.

[86] A. Sebastian, M. Le Gallo, and D. Krebs, "Crystal growth within a phase change memory cell," *Nature Commun.*, vol. 5, p. 4314, Jul. 2014, doi: 10.1038/ncomms5314.