Test and Reliability of Emerging Non-Volatile Memories

[1]Said Hamdioui, [1]Peyman Pouyan, [2]Huawei Li, [2]Ying Wang, [3]Arijit Raychowdhur, [3]Insik Yoon

[1]Laboratory of Computer Engineering, Delft University of Technology, the Netherlands

[2]Institute of Computing Technology, Chinese Academic of Sciences

[3]Laboratory of Computer Engineering, Georgia Institute of Technology

Email: {s.hamdioui, p.pouyan}@tudelft.nl {lihuawei, wangying2009}@ict.ac.cn {arijit.raychowdhury, iyoon}@ece.gatech.edu

Abstract—The search for alternative memory technologies has attracted significant attention toward emerging non-volatile memories. Among them, STT-MRAM, PCM, RRAM have shown promising characteristic to gain a position inside the memory hierarchy of computing platforms, and even enable new computing paradigms. However like any other emerging technology these devices are affected by concerns to be resolved before they could become a mainstream. This paper reviews the main reliability and testability challenges of aforementioned emerging non-volatile memories and highlights the main future considerations toward them.

Index Terms—STT-MRAM, PCM, RRAM, Test, Reliability

I. Introduction

Semiconductor memories have been evolving over time; e.g., SRAM for primary memory (cache), DRAM for secondary (main memory), and Flash for mass-storage. Recent/emerging applications (such as big data applications and internet-of-things) are extremely demanding; not only in terms of computing power, but also in terms of storage. They have additional requirements for memory systems (e.g., higher bandwidth, higher density, lower power, sustainable scaling, lower cost, lower latency) [1]; and even the performance of computer systems are heavily dependent on the the characteristics of the memory subsystem. Traditional memories are not able to satisfy all of these requirements. Moreover, they are facing major challenges such as the limited scaling and increased manufacturing cost for smaller nodes [2]. Therefore, a lot of effort is put on searching and developing new memory alternatives. The main research focuses on the non-volatile memory technologies as alternative technologies to satisfy the high demands of future applications. Among the most relevant emerging memory technologies today are the ones built with resistive devices; the International Technology Road map of Semiconductors (ITRS) in its 2015 report identified the Spin Transfer Torque MRAM (STT-MRAM), Phase Change Memory (PCM), and Resistive RAM (RRAM) as the most promising memory technologies with potential for scaling and commercialization of non-volatile RAM to and beyond the 16nm generation. Today both STT-MRAM and PCM can be considered as "'prototypical", while RRAM is still emerging. For instance Globalfoundries is planning to manufacture 22nm embedded memory using STT-MRAM [3], Samsung has already PCM products [4], Scandisk/Toshiba and Crossbar have presented fabricated RRAM prototype chips [5] [6]. All of these NVM memory technologies use two-terminal structure devices to construct the memory cell array, and can be easily integrated in the back end of line (BEOL) of CMOS process. They are also compatible with a crossbar array structure where memory elements are built at the crossing points of horizontal and vertical access lines; hence, enabling high density storage especially when considering multiple stacked layers . Obviously, providing appropriate product quality and reliability to the market is a key enabler for the successful commercialization of such memories [7].

Although people believe that STT-MRAM and PCM are to be commercialized soon (if not already at a small scale) and that RRAM will come later, limited work has been done on the way to test these memories in order to guarantee the outgoing product quality and reliability. This paper provides and overview on exiting test schemes for these three memories and highlights the challenges that have still to be worked out.

The rest of the paper is organized as follows. Section II address test and reliability of STT-MRAMs. It first reviews the basics of STT-MRAM; then it covers their defect and fault analysis; and finally it discusses their testing. Section III covers test and reliability of PCMs; it provides first the basic memory operations; then it addresses their testability and thereafter their reliability. Section IV does the same as Section III, but then for RRAMs. Section V highlights major challenges in the discussed emerging non-volatile memories. Finally Section VI presents some conclusion of the work.

II. Test and Reliability of STT-MRAMs

A. STT-MRAM Basics

Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM) is an emerging memory technology which exhibits non-volatility, high density and nanosecond read and write times. These attributes of STT-MRAM make it suitable for last level embedded caches. However, the defects and corresponding fault models of STT-MRAM are not as extensively explored as in SRAM and therefore, there is a growing need for defect and fault analysis. Moreover, stochastic retention failure of STT-MRAM imposes a large burden in testing time. Conventional test schemes for retention of STT-MRAM need to be optimized for testing a large-size embedded STT-MRAM

array. Here we present a review of the different defect and fault mechanisms as well as a BIST architecture and circuit to reduce testing time in characterization and manufacturing tests for retention. We address the effect of resistive and capacitive defects and identify retention test setup for measuring worst case retention.

A STT-MRAM cell is composed of one access transistor and one magnetic tunnel junction (MTJ) that stores a bit of information as shown in Fig 1.(a) [8]. An MTJ has two ferromagnetic layers (CoFeB based), which are called fixed and free layer and they are separated by a thin insulator layer(MgO based). The magnetic moment in fixed layer is 'fixed' into one direction and the direction of magnetic moment in free layer can be changed depending on difference and polarity of potential across an MTJ. When a potential difference is applied across MTJ, spin-polarized current passes through an MTJ and it attempts to polarize the current in its preferred direction of magnetic moment. The angular momentum of the electrons in free layer creates a torque that causes a flip in the direction of magnetization inside the free layer of MTJ. Depending on the direction of magnetization in the free layer, resistance of MTJ is changed. As shown in Fig 1, when the direction of magnetization in free layer is anti parallel to the magnetization of fixed layer, MTJ has high resistance and when they are in parallel, MTJ has low resistance. Bit 1 and 0 are mapped to the two cases: when MTJ has high resistance and low resistance respectively. The bias conditions on bit line, source line and word line for write and read operation are shown in Fig 1.(b)-(d). The write operation is bidirectional, where either the bitline (BL) or source line (SL) is pulled high and the other one is pulled low depending on the polarity of the write operation. In case of writing 1, the bit line is vdd and source line is ground. When word line is asserted, write current flows from source line to bit line and MTJ will be in anti-parallel state. The bias condition is set to the opposite when writing 0 to a cell and the current from bit line to source line sets MTJ to be in parallel state. The read operation is unidirectional with word line driven to $vdd/2$ and a pre-charged BL voltage discharges through the cell. Depending on the resistance of MTJ, discharging voltage at source line is sensed at sense amplifier. STT-MRAM is non-volatile since a bit is stored in an MTJ as resistance and it is determined by magnetization of free layer. The MTJ can either be an In-plane MTJ (I-MTJ) with magnetic anisotropy in plane due to shape anisotropy or a Perpendicular plane MTJ (P-MTJ) where magnetic anisotropy is aligned out of plane, independent of the shape of the free layer[9]. The relative merits and demerits of the two structures are being extensively studied [9][10][11][12] .

STT-MRAM arrays are expected to suffer from read and write failures which are induced by electrical defects and process variations. In [8][13], the types of resistive, capacitive and coupling defects are identified. Their manifestation as read and write faults as well as fault activation patterns are analyzed. Apart from read and write faults, STT-MRAMs can also suffer from retention failures, a bit-flip in a cell caused by thermal noise [14, 15]. Since retention time is exponentially
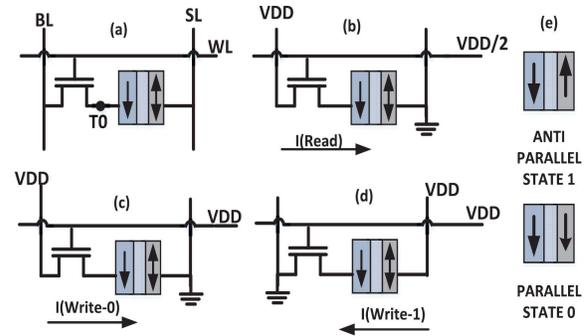

Fig. 1: Basic STT-MRAM cell (a) 1T-1MTJ representation (b) bias condition for read (c) write 0 bias condition (d) write 1 bias condition (e) states in a MTJ due to orientation of magnetic moments

proportional to the stored energy (thermal stability), conventional test method for retention described in [14] measures thermal stability of a cell by applying weak write current to a cell. However, it results in prohibitively large test time. In the following of this paper we review a Memory Built In Self-Test (MBIST) architecture from [16] that can detect the retention failures in a time-efficient manner. It is an efficient MBIST architecture that can perform in-situ read, write and retention (stochastic test) tests on STT-MRAM arrays.

B. Defects and fault analysis

[13][8] have identified faults during read and write and how resistive and capacitive defects induce faults. Table I summarizes fault models and how they contribute to read/write failures.

Fault Model	Affects	Key Cause
Transition Fault (TF)	WR	Relative Weak WR current due to stray resistive paths
Coupling Fault(CF)	WR	Neighboring cells switching
Stuck At Fault(SF)	WR	T0 , WL stuck at VDD or GND
Incorrect Read Fault (IRF)	RD	Current miscorrelation due to defects affecting WL,BL
Read Disturb Fault (RDF)	RD	Electrical disturbance at T0 node due to larger than normal RD current

Fig. 2: Table I: Defects induced faults[16]

1) Resistive defects: In this section, we discuss the role of resistive defects in 2 by 2 cell array explored in [8][16]. As shown in Fig.3, the resistive defects are associated with resistive shorts between the node of victim cell and aggressor cells. In the figure, 13 possible resistive defects can affect read/write of the cell.

For example, a case of having $RS_{BL0-SL1}$ short between BL0 and SL1 in Fig.3(a) results in write 0 fault in cell-0 because $RS_{BL0-SL1}$ prohibits BL0 to reach full VDD.

As a result, weak write current across cell-0 causes a transition fault to 0. A read failure can happen due to $RS_{WL0-WL1}$. In case of reading cell-0, word line 0 is weakly asserted due to the resistive short between wordline 0 and 1. If cell-1 contains 0, a change of incorrect read fault can happen because of slower discharge of bitline 0. [8] analyzed how each resistive short shown in Fig.3 causes faults in either read or write operation.
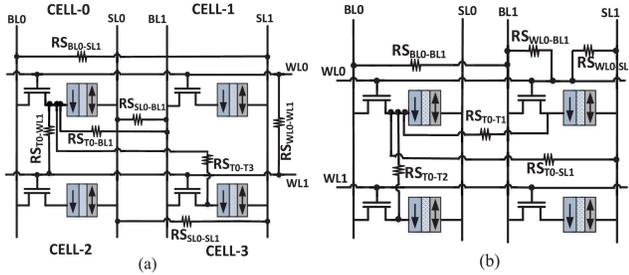

Fig. 3: (a) and (b) Inter-cell resistive defects[8]

2) Capacitive defects: It has been noted in [8][16] that capacitive defects, especially bridging capacitance plays a weak role in causing failures. The readers are directed to [16] for more discussions on this.

C. Retention and Thermal Stability Tests and MBIST Architecture

In order to measure retention, authors in [14] proposed a possible test methodology based on the thermal activation model. In STT-MRAM, retention time is defined as the time it takes for a cell to flip, a stochastic phenomenon, caused by thermal noise [12]. However, since retention testing needs to be done with a weak current and stochastic measurements need to be collected, the total test time increases significantly. In order to solve retention test time problem, [16] proposed a new retention test scheme that perform in-situ, statistical, retention testing of large STT-MRAM arrays. From the retention BIST algorithm [14], a weak write current is applied row by row and the row is read right after to obtain P_{sw}. The major drawbacks from this scheme are; (1) The retention test time increases linearly when the row size of an array increases.

(2) The retention tests have to be carried out in an linear region where P_{SW} is very low. Since P_{sw} is very low in the region that retention test is performed, a bit flip will not happen for most of the iterations; which means most of the read operations after applying current are not necessary.

These two problems are main bottlenecks for improving speed of retention test. The retention test scheme from [16] reduces retention time significantly by:
1. Applying weak write current to multiple rows
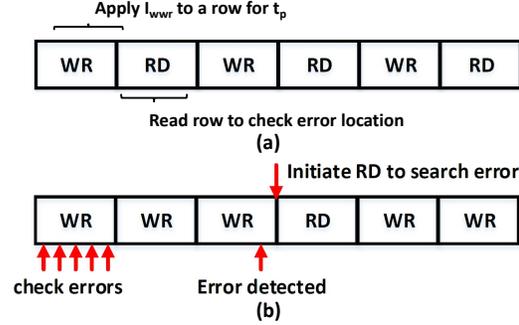2. Avoiding search (reading rows) when error is not detected

Fig. 4: (a) conventional test scheme (b) proposed test scheme

By testing multiple rows in a column at the same time and searching for errors after error detection, retention time testing reduces significantly. As Fig.4 shows, the new test scheme concurrently runs error detection while applying weak write current to multiple rows. Read operation happens only when error is detected within the rows under test.

The retention test is divided into two processes, (1) Error Detection (ED) and (2) Error Search (ES).

1) Error Detection (ED): The ED architecture detects an error by observing the change in current flowing through a cell since a bit flip in a cell changes resistance ac cross a cell and it results in current change. As shown in Fig. 5, MTJ resistances in multiple cells in a column are connected in parallel when corresponding word-lines turn on simultaneously. When I_{WWR} causes a bit flip in a cell, the current at source line (I_{SL}) changes due to the resistance change, which shows that at least one error is detected within a column [16].
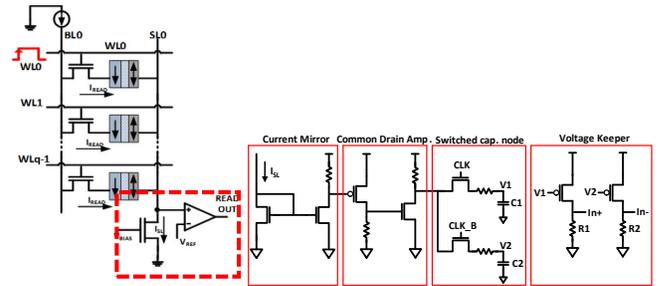

Fig. 5: Error Detection circuit for a column with 16 rows[16]

2) Error Search: Once error is detected from ED stage, The location of error is searched within activated rows in order to obtain P_{sw} and thermal stability of cells. [16] presented exhaustive search and temporal locality search algorithms and interested readers are pointed to [16] for more discussions.

III. TEST AND RELIABILITY OF PCMS

A. Primer on Phase Change Memory

Phase Change Memory (PCM) is one of the non-volatile memory (NVM) technologies that are most likely to be deployed in commercial products. Recently, prototypes and
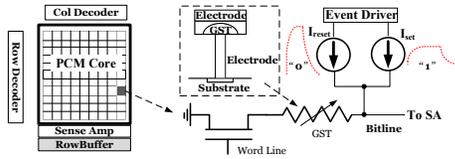
Fig. 6: Hierarchical organization of typical PCM array.

off-the-shelf of PCM modules have already been issued by companies including Samsung and Numonyx [4, 17, 18]. With its superior performance, storage density and non-volatility, PCM could provide a promising DRAM alternative to build the futuristic scalable memory systems.

Typically, Phase-change Random Access Memory (PRAM) has the same hierarchical organization as DRAM. As shown in Fig 6, a PRAM bank contains multiple arrays, and each array consists of a large number of cells. As illustrated in Fig 6, a typical PCM cell contains a layer of glass chalcogenide materials such as $Ge_2Sb_2Te_5$ (GST) as a variable transistor between two electrodes. The chalcogenide can be either in an amorphous state or a crystalline state, each of the states represents a logic value stored in the cell. When a read command switches on word line, the read current can pass the resistor and generate a voltage fluctuation in the bitline that is sensed by the Sense Amplifier (SA) to drive the output. For write operations, the cells state can be switched when the electrodes are heated up to a high temperature by a programming current. The heating current I_{reset} and I_{set} are featured with different patterns, and they can generate Joule heat that transforms chalcogenide into high or low resistance depending on the amplitudes and duration of the programming currents. For RESET operation, I_{reset} has to produce enough heat to crystallize the chalcogenide, so it has a sufficiently high magnitude. In contrast, I_{set} has a lower peak amplitude, but lasts longer to heat the GST to the threshold of transition temperature. Both of the Set/Reset operation consumes a great amount of write power, and the induced programming current can physically degrade the endurance of phase change material, limiting the cycle lifetime of PCM cells [19]. This critical state-change behavior and how to mitigate the overhead of write operations have been the focus of prior research study and design efforts from both academia and industry to improve the reliability, performance and power characteristics of PCM devices.

B. Testing of PCMs

Due to its special materials, device structure, and operation mechanisms, PCM has its specific failure modes. A comprehensive introduction of its failure modes and fault models of PCM can be found in [20]. Typical fault models of PCM include: 1) disturb faults such as proximity (PD), read recovery (RRD), read (RD) and false write (FWR), 2) program faults (PF), 3) stuck-at SET (SS) and stuck-at RESET (SR) faults, 4) false write (FWR) faults, 5) transition faults (TF), and so on.

New March algorithms have been explored in recent years to cover the specific failure modes induced faults of PCM. In [20], a March algorithm called March-PCM is introduced to detect PCM specific faults in addition to common faults. In [21], March-PCM^P is presented to detect additional faults named weak transition faults and write destructive faults caused by the parasitic capacitance and resistance defects in stand-alone PCM cells. In [22], PCM errors are characterized into four types: program interference faults, read faults, and write 1 and write 0 faults, and basic March algorithms are utilized to classify and detect them. In [23], a testing scheme based on "sneak-path sensing" is proposed to efficiently detect faults in crossbar-based NVMs, so that a group of memory cells can be tested simultaneously, thereby reducing testing time. DFT supports are needed for sneak-path testing. An enhanced sneak path test algorithm is proposed in [24] to extend the fault coverage capability of RD, FWR, RRD, and PF faults.

C. Reliability of PCMs

One of the major reliability concerns of PCM is the lifetime reliability, i.e., the problem with wearout. Repetitive writing causes a PCM cell ware, eventually leading to a hard fault which corrupts the data in it. Researchers have presented techniques on write traffic reduction, wear leveling, and salvaging [25] to overcome the limitations of write endurance. While write traffic reduction and wear leveling strive to avoid hard faults in PCM cells, the purpose of salvaging is to correct errors during normal operations after hard faults have occurred.

Write traffic reduction techniques include data comparison write (DCW), data inversion (DI), approximate write (AW), and coding techniques. DCW reduces write activities by only writing the cells whose current value is different from the value to be written [26, 27]. DI extends DCW to achieve fewer bit writes by calculating the hamming distance between the old data and the new one to determine whether to write the new one or its inverted counterpart [26], [28]. AW sacrifices the data integrity for suppressing write operations, which can be utilized in applications which exhibit the characteristic of inherent error tolerance and dont require absolute correctness of the outputs [29].

Encoding the data before write can potentially reduce the write traffics in PCM arrays [30]. FlipMin is such a technique that minimizes the number of flip-bits by using coset coding. Other coding formats are aware of the write symmetry in cell programming. As depicted in Fig 6, the RESET current is the deciding factor to the lifetime of PCM when compared to SET. There is a plenty of work aiming at minimizing the frequencies of RESET operations in PCM via an asymmetry-aware data encoding method [31]. For video applications where PCM is used as main memory, [32] proposed inter-block differential data encoding and inter-frame multiple experts to reduce write operations.

An interesting work to reveal that limited write endurance of PCM incurs a potential security threat is presented in [33].
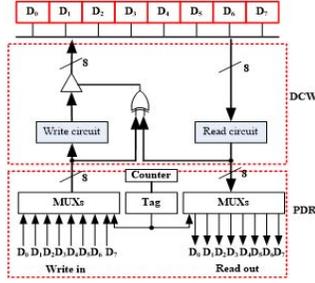
Fig. 7: Periodical data reversion after DCW [34].

From the standpoint of attackers, random stream attacks are performed for PCM used in video applications. The attacks cause extremely high write traffic or worsened lifetime, which cannot be handled by the existing write traffic reduction techniques.

Wear leveling techniques attempt to distribute write operations uniformly across the PCM cells. Related work includes row shifting & segment swapping [35], fine grained wear leveling [36] and start-gap for address randomization [37]. Besides, in the hybrid DRAM and PCM main memory [38], hot pages from PCM are swapped into DRAM to achieve wear leveling.

In [34], periodical data reversion (PDR) is proposed for wear leveling and error tolerance (ET) to enhance lifetime of PCM-based image buffer after DCW. DCW first eliminates inherent redundant writes by taking advantage of temporal redundancy between successive video frames. PDR exchanges data write locations for wear leveling, while ET extends effective lifetime with graceful degradation of video quality and compression ratio. Wear leveling using PDR needs a little extra hardware overhead, as shown in Fig 7, but incurs no quality degradation. In certain cases where strict correctness is desired, wear leveling can be used. ET, as an abnormal way, enhances effective lifetime without additional hardware overhead. It may be used in certain application scenarios where a slight degradation of quality or compression ratio is acceptable.

Salvaging techniques involve ECC, ECP [39], SAFER [40] and dynamically replicated pages [41]. Usually salvaging incurs reduced effective memory capacity or performance degradation. Traditional ECC is costly. ECP restores faulty cells by storing the error location and the corresponding correct value [39]. SAFER reduces area overhead for error correction by utilizing the fact that cells with stuck-at-faults are still readable [40]. In [42], wear leveling and salvaging techniques are integrated together to improve the lifetime of PCM main memory.

PCM also suffers from soft errors due to resistance drift or write disturb faults. A detailed discussion on the cause of soft errors, and solutions to detect and correct soft errors can be found in [43]. In [44], the authors investigated the write-induced IR-drop problem and its effect on 3-D-stacked PCMs. The IR-drop violation poses a serious threat that enforces a strict guard band of requesting concurrence, and consequently reduces the write throughput. The authors then proposed a power supply integrity conscious write scheduler for 3-D Die-Stacked PCMs to improve the write performance while maintain reliable write operations.

IV. TEST AND RELIABILITY OF RRAMS

A. Primer on Resistive Random Access Memory

Resistive Random Access Memory (RRAM) is a NVM type that functions based on a change of resistance value on a Metal-Insulator-Metal (MIM) structure because of ion migration/filament creation inside the structure along with some redox processes including electrode/insulator material [45]. According to ITRS, RRAM has been categorized to four main types based on filamentary functioning and switching property [2], where two of the most important subcategories are Oxide-RRAM (OxRAM) and Conductive Bridge RAM (CBRAM). OxRAM and CBRAM's switching principle are based on creation and rupture a conductive filament in the insulator layer between the two electrodes. In OxRAM the filament is constructed by oxygen vacancies while the CBRAM's filament is generally formed by metal atoms.

Fig 8.a shows an example of filamentary RRAM, in its two resistance state modes. Initially and just after the manufacturing a forming process is applied to the device which forms a filament between top and bottom electrodes without connecting them to each other. Later applying a positive voltage at the top electrode will extend the Conductive Filament (CF) formation between the metal contacts and reduce the resistance toward the Low Resistance State (LRS) mode, this is called a SET Process. Applying a voltage with opposite polarity would reverse the ion migration process and will rupture the CF toward the High Resistance State (HRS) mode, this is called a RESET Process.

To construct a memory cell, the RRAM can be exploited alone (1R) or with a two terminal selector device (1S1R) to construct a crosspoint array [46]. Alternatively the RRAM based memory cell can be built up with one transistor forming a 1T1R array architecture. The 1T1R architecture has been widely researched as it removes the sneak path problem in 1R array architecture [23]. Fig 8.b presents an example of 2by2 1T1R architecture, where the RRAM is connected to the Bitline (BL) and a transistor which is connected to Wordline (WL) and Sourceline (SL). Employing appropriate voltages at these terminals (WL,BL,SL) can SET/RESET the device and also makes it possible to read the state of RRAM through the Sense-Amplifier (SA) [47].

B. Testing of RRAMs

Similarly to other emerging non-volatile memories, RRAM is susceptible to defects due to imperfect manufacturing conditions, therefore appropriate test mechanisms are required to detect such failures [48] [49]. However, in addition to conventional faults in RAMs, RRAM due to its specific physical
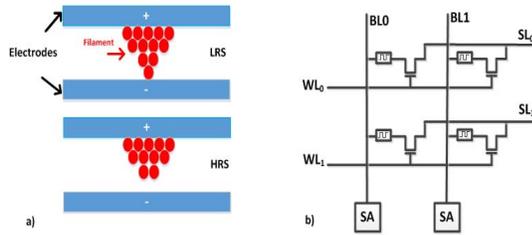
174

Fig. 8: a)Filamentary RRAM, b)1T1R architecture

Fig. 9: Short Write time-based DFT

construction and switching mechanism can be affected by unique faults such as over-forming and undefined state ones [48] [50]. These special faults have caused the test engineers to modify the traditional March test algorithms and also have signified the importance of utilizing new Design for Testability (DFT) techniques to cover all fault types [23] [51] [50].

Specific March test algorithms have been developed in order to perform RRAM memory testing and to detect their unique faults in [48] [23] [51]. Regarding these, [48] proposes a modified March C algorithm to diagnose the over-formed faulty cells from the good cells by adding two consecutive read '1' operations and removing the initial write step from the original algorithm. In [23] the authors propose a March test based on sneak path current in the crossbar type RRAM memory that can test multiple cells at once. In this approach the memory peripherals are adjusted to function normally at operational mode and to introduce sneak path during the test mode. Any defect in the memory cell can affect the sneak current and help to detect the faulty cells. Finally, [51] presents an extended March approach to find the defective RRAM memory cell constructed with a transistor (1T1R). The March 1T1R can properly detect the cell faults caused by a defective transistor in the RRAM memory cell.

Due to analog nature of the RRAM device, its unique fault of undefined state cannot be detected by any conventional March algorithm [52]. This is due to fact that March tests deal only with fixed, predetermined patterns of logic values. This fault however causes a random logic value to be read from the defective RRAM cell. Regarding this, [50] have introduced a DFT scheme to detect such specific RRAM fault. This technique is based on either Short Write time-based DFT or Low write voltage-based one. In both cases the target is to generate a write pulse which is not big enough to change the status of the good cell but to drift the faulty cell from the undefined region to the incorrect known state and to able to detect it properly in the following read step. Fig 9 shows an example for a Short Write time-based DFT [50].

C. Reliability of RRAMs

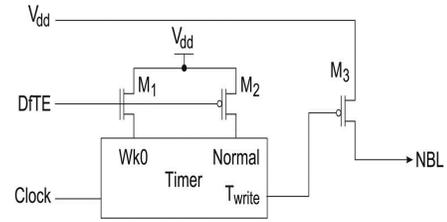Like any emerging technology, RRAM devices are also affected by some reliability concerns due to immature manufac-

turing flow. This section reviews three of the main reliability issues in RRAM devices.

Parametric Variation in the high/low resistance values is a key reliability challenge in the design of RRAM circuits [53] [54]. These fluctuations are categorized into two types: 1-Cycle to cycle variation happening in each switching cycle 2-Device to device variation where the resistance value differs in each fabricated device. RRAM variability arises from deviations in the conductive filament. This is mainly due to stochastic nature of ion migration. For instance the shape, size or the gap distance in the filament may vary and impose resistance fluctuations . It is worth to note that the RRAM variability can be affected by operating conditions such as temperature and voltage [55].

Endurance Degradation is the second important reliability concern in RRAM devices which causes the limited number of write cycles [56]. This mechanism originates from SET-RESET switching properties in the device [57]. The SET process is correlated with a soft breakdown of the resistive switching layer. The oxygen ions and oxygen vacancies are generated by the electric field. Then, the oxygen ions get drifted to the anode and the existing oxygen vacancies construct a conducting filament, and the resistance value switches from high to low. Degradation mechanisms cause that the conducting path to become larger than the nominal one and the path ruptures more difficult. This type of degradation is called as Over-SET [58]. Other type of endurance degradation can occur in the the RESET process. In this phase the recombination of oxygen ions and oxygen vacancies will rupture the filament and cause a switch from low to high resistance. But when the filament gap is larger than nominal the conductive path formation becomes more difficult and Over-Reset degradation occurs [58]. This mechanism depends on different parameters, among others, the environment temperature and switching speed [59].

Random Telegraph Noise (RTN) is another reliability concern in RRAM devices. It causes current fluctuations at high and low resistance values due to activation/deactivation of the electron traps inside the filament [60]. The current variation by RTN can induce read instabilities and reduces the memory read window in RRAM memories if enough consideration is not taken care.

175

These reliability challenges in the design of RRAM based memories have prompted the techniques to improve their dependability. These approaches can be based on enhancing the quality of device through material improvement [45], by circuit-based methods such as optimizing the RRAM programming operation to increase the number of write cycles [57], to design adaptive readout circuitry to better tolerate RRAM resistance variability [61] and to innovate new reconfiguring techniques to extend the RRAM lifetime [62].

V. FUTURE CHALLENGES

Every type of emerging non-volatile memory have unique features and therefore can have various applications in the memory hierarchy. To be considered as a competent rival for conventional memories such as SRAM, DRAM and Flash, they all need to further reduce their cost per bit and improve their reliability and testability characteristics [63]. The following briefly discusses the main future challenges with each emerging non-volatile memory type.

STT-MRAM in spite of their promise, provide a number of fundamental challenges in technology enablement. The key problem in STT-MRAM bit-cells is the large stochasticity in the process of write. Since the write process is thermally driven, the same bit cell requires a variable amount of time to complete the write process; thereby causing a long tail of the write current and write time distributions. This creates serious challenges in architecture design, where a significant overhead needs to be spent to successfully write into an array. This leads to corresponding challenges in array testing as well, where the test time increases or good bits are marked as faulty. The work described in the paper addresses this key issue, but more work needs to be done to make the test-time of STT-MRAM comparable to SRAM.

Smarter test algorithms, parallel test of sub-arrays, as well as statistical processing of test-data can be key enablers to identifying weak bits and providing protection via ECC and redundancy. Coupled with the increasing test-time because of stochasticity, STT-MRAM cells also suffer from time dependent dielectric breakdown (TDDB). As a large write current and hence electric field is required to write the bit cell, the repeated stress on the oxide, leads to eventual breakdown thereby compromising the cell's read and write properties. To achieve high reliability under TDDB and to detect bit cells with compromised dielectrics, we need test strategies that can detect marginal bit cells during manufacturing test. Although research continues in the earnest to enable test strategies in shortening the test time of STT-MRAM as well as designing innovative BIST structures, we need concerted effort from the industry and academia to identify unique STT-MRAM test challenges and provide solutions to make this a viable technology.

PCM currently has two major design challenges to address in practical systems: write endurance and resistance drift. From system-perspective, PCM devices are positioned and expected to be used in high-density main memory or storage system. Firstly, when replacing the current DRAM with PCM in main memory system, the write endurance issue is still the greatest challenge to address, as a result of the high access frequency of working-set memory. How to enable inexpensive protection schemes for PCM, avoid their negative impacts on memory bandwidth and minimize the capacity losses are very important to the main memory, whose primary goal is to provide high-bandwidth, stable and low latency contemporary storage for the emerging memory-intensive workloads in big data era. Secondly, as the storage or secondary memory alternative to FLASH, multi-level cell PCM is superior in both performance and write endurance, but it is required to maintain the cell state for a much longer period as a permanent storage device than in main memory. Thus the problem of resistance drift is more pronounced in storage than in DRAM. It is expected to consider and use drift tolerant coding, detection or even data refresh schemes other than the conventional error detection and correction methods in solid-state storage devices. Finally, irrespective of the application, the circuit-level challenge to PCM scaling is the increasing variability when feature size keeps shrinking. Parametric fluctuation and variations make PCM more susceptible to intensive write behavior, undesirable read/write noise and drift. This imposes the necessity to design effective MBIST or even Memory Build in Self-Repair (MBISR) architectures which support concurrent testing for PCM arrays under the power and heat constraints, while covering the critical failure modes of PCM. In addition, seeking cross-layer fault tolerant and treating device failure as the normal case at design time are likely to bring newer opportunities in futuristic PCM based systems.

RRAM's main challenge is the variability of its switching parameters due to the stochastic nature of the ion migration/filament creation [63]. This variability not only enforces device by device fluctuations but also results in cycle to cycle deviations, making it crucial to properly read the device state. Such a concern imposes the necessity to design adaptive sensing circuits to correctly identify the RRAM resistance value. Another solution is to utilize write verification steps inside the memory, so the state of RRAM can be set precisely for the proximate read cycles [64].

Scaling is the next challenge with RRAM devices. Although RRAM can be scaled down to few nm sizes due to its atomic switching principle, special considerations are needed for highly scaled devices. The RRAM's data retention capabilities becomes weak when its filament is too thin [65] and this can result in reliability concerns in future nano-scale RRAMs.

Finally, testing of RRAM devices can be a considerable challenge for future high capacity RRAM memories. Till recently just a few prototypical RRAM chips have been fabricated and tested in the lab [48]. Their testing have shown new fault mechanisms inside RRAM memories [48]; faults which cannot be completely detected with conventional test approaches. Therefore there is a need for research and

176

investigation of new design for testability approaches for RRAM based memories. These techniques should be fast and efficient in covering all potential faults in highly scaled RRAM memories.

VI. CONCLUSION

This paper has briefly reviewed testability and reliability aspects of three emerging but promising non-volatile memories; these are STT-MRAM, PCM and RRAM. Interesting enough is that although all these three memories have some common features (such as being two terminal resistive-based storage devices), their failure mechanism and therefore they way they have to be tested are quite different; this is due to difference in their switching mechanisms and cell structures, which impact their reliability and test challenges. Hence, not only that traditional methods for memory testing (e.g. SRAMs) cannot guarantee the required outgoing product quality for such memories, but also each of them requires specific approaches and DFT to deal with their distinctive and/or faults. Moreover reliability and testability enhancement techniques are also required to improve their dependability. Also industrial designs and data are still missing in the community, making it hard to make a realistic judgment of the published solutions and their weaknesses. It is expected that this field will get more attention especially now that some of these emerging memories are getting closer and closer to the commercialization. Manufacturing test is the latest step/chance to satisfy the customer requirements in terms of quality and reliability; hence the importance of high quality, but cheap test solutions.

REFERENCES

[1] M. Pavlovic, Y. Etsion, and A. Ramirez, "On the memory system requirements of future scientific applications: Four case-studies," in *IEEE International Symposium on Workload Characterization (IISWC)*, 2011.
[2] "http://www.itrs.net/," *International Technology Roadmap for Semiconductors*, 2015.
[3] "https://www.globalfoundries.com," *Global Foundries*, 2016.
[4] H. Chung, B. H. Jeong, et al., "A 58nm 1.8V 1Gb PRAM with 6.4MB/s program BW," *ISSCC*, pp. 500–502, 2011.
[5] T. yi Liu and et al., "A 130.7-mm2 2-Layer 32-Gb ReRAM Memory Device in 24-nm Technology," *IEEE Solid States Circuits*, vol. 49, no. 1, pp. 140–153, September 2013.
[6] "https://www.crossbar-inc.com," *Crossbar*, 2017.
[7] E. I. Vatajelu, P. Prinetto, M. Taouil, and S. Hamdioui, "Challenges and Solutions in Emerging Memory Testing," *IEEE Transactions on Emerging Topics in Computing*, vol. PP, no. 99, pp. 1–1, April 2017.
[8] A. Chintaluri, H. Naeimi, S. Natarajan, and A. Raychowdhury, "Analysis of defects and variations in embedded spin transfer torque (stt) mram arrays," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. PP, no. 99, pp. 1–11, 2016.
[9] S. A. Wolf, J. Lu, M. R. Stan, E. Chen, and D. M. Treger, "The promise of nanomagnetics and spintronics for future logic and universal memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2155–2168, 2010.
[10] G. Jan, L. Thomas, S. Le, Y.-j. Lee, H. Liu, J. Zhu, 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, "Demonstration of fully functional 8Mb perpendicular STT-MRAM chips with sub-5ns writing for non-volatile embedded memories," *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, vol. 093008, no. 2012, pp. 1–2, 2014.
[11] H. Yoda, S. Fujita, N. Shimomura, E. Kitagawa, K. Abe, K. Nomura, H. Noguchi, and J. Ito, "Progress of STT-MRAM technology and the effect on normally-off computing systems," *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 259–262, 2012.
[12] A. V. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R. S. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. H. Butler, P. B. Visscher, D. Lottis, E. Chen, V. Nikitin, and M. Krounbi, "Basic principles of STT-MRAM cell operation in memory arrays," *Journal of Physics D: Applied Physics*, vol. 46, no. 13, p. 139601, 2013.
[13] A. Chintaluri, A. Parihar, S. Natarajan, H. Naeimi, and A. Raychowdhury, "A Model Study of Defects and Faults in Embedded Spin Transfer Torque (STT) MRAM Arrays," *2015 IEEE 24th Asian Test Symposium (ATS)*, vol. 1, no. c, pp. 187–192, 2015.
[14] H. Naeimi, C. Augustine, A. Raychowdhury, S.-l. Lu, and J. Tschanz, *Intel Technology Journal, STTRAM Scaling and Retention Failure*, 2013, vol. 17, no. 1.
[15] R. Heindl, W. H. Rippard, S. E. Russek, M. R. Pufall, and A. B. Kos, "Validity of the thermal activation model for spin-transfer torque switching in magnetic tunnel junctions," *Journal of Applied Physics*, vol. 109, no. 7, 2011.
[16] I. Yoon, A. Chintaluri, and A. Raychowdhury, ""EMACS: Efficient MBIST Architecture for Test and Characterization of STT-MRAM Arrays," *International Test Conference*, 2016.
[17] G. Servalli, "A 45nm Generation Phase Change Memory Technology," *IEDM*, pp. 113–116, 2009.
[18] G. W. Burr, M. J. Brightsky, et al., "Recent progress in phase-change memory technology," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 2, pp. 146–162, 2016.
[19] K. Kim and S. J. Ahn, "Reliability investigation for manufacturable high density PRAM," *IRPS*, pp. 157–162, 2005.
[20] M. G. Mohammad, "Fault model and test procedure for phase change memory," *IET Comput Digital Tech*, vol. 5, no. 4, pp. 263–270, 2011.
[21] X. Pan, X. Cui, et al., "Modeling and test for parasitic resistance and capacitance defects in PCM," *NVMTS*, pp. 73–76, 2012.
[22] Z. Zhang, W. Xiao, et al., "Memory module-level testing and error behaviors for phase change memory," *ICCD*, pp. 358–363, 2012.
[23] S. Kannan, J. Rajendran, et al., "Sneak path testing of crossbar-based non-volatile random access memories," *IEEE Trans. Nanotechnol.*, vol. 12, no. 3, pp. 413–426, 2013.
[24] X. Cui, Z. Cheng, et al., "A snake addressing scheme for phase change memory testing," *Sci China Inf Sci*, vol. 59, no. 10, p. 102401, 2016.
[25] C. Xue, Y. Zhang, et al., "Emerging non-volatile memories: opportunities and challenges," *CODES+ISSS*, pp. 325–334, 2011.
[26] Y. Joo, D. Niu, et al., "Energy- and endurance-aware design of phase change memory caches," *DATE*, pp. 136–141, 2010.
[27] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "A durable and energy efficient main memory using phase change memory technology," *ISCA*, pp. 14–23, 2009.
[28] S. Cho and H. Lee, "Flip-N-Write: a simple deterministic technique to improve PRAM write performance, energy and endurance," *MICRO*, pp. 347–357, 2009.
[29] Y. Fang, H. Li, and X. Li, "SoftPCM: Enhancing Energy Efficiency and Lifetime of Phase Change Memory in Video Applications via Approximate Write," *ATS*, pp. 131–136, 2012.
[30] A. N. Jacobvitz, R. Calderbank, and D. J. Sorin, "Coset coding to extend the lifetime of memory," *HPCA*, pp. 222–233, 2013.
[31] A. Mirhoseini, M. Potkonjak, and F. Koushanfar, "Coding-based energy minimization for phasechange memory," *DAC*, pp. 68–76, 2012.
[32] S. Kwon, S. Yoo, S. Lee, and J. Park, "Optimizing video application design for phase-change RAM-based main memory," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 11, pp. 2011–2019, Nov. 2012.
[33] Y. Fang, H. Li, X. Li, "RSAK: Random Stream Attack for Phase Change Memory in Video Applications," *VTS*, pp. Paper 10B–3, 2013.
[34] Y. Fang, H. Li, and X. Lis, "Lifetime enhancement techniques for PCM-based image buffer in multimedia applications," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 6, pp. 1450–1455, 2014.
[35] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting phase change memory as a scalable DRAM alternative," *ISCA*, pp. 2–13, 2009.
[36] M. K. Qureshi, V. Srinivasan, and J. A. Rivers, "Scalable high performance main memory system using phase-change memory technology," *ISCA*, pp. 24–33, 2009.
[37] M. K. Qureshi, J. Karidis, and M. Franceschini, "Enhancing lifetime and security of PCM-based main memory with start-gap wear leveling," *MICRO*, pp. 14–23, 2009.
[38] G. Dhiman, R. Ayoub, and T. Rosing, "PDRAM: a hybrid PRAM and

DRAM main memory system," *DAC*, pp. 664–669, 2009.

[39] S. Schechter, G. H. Loh, K. Strauss, and D. Burger, "Use ECP, not ECC, for hard failures in resistive memories," *ISCA*, 2010.

[40] N. H. Seong, D. H. Woo, V. Srinivasan, J. A. Rivers, and H.-H. S. Lee, "SAFER: Stuck-At-Fault Error Recovery for Memories," *MICRO*, pp. 115–224, 2010.

[41] E. Ipek, J. Condit, E. Nightingale, D. Burger, and T. Moscibroda, "Dynamically replicated memory: building reliable systems from nanoscale resistive memories," *ASPLOS*, pp. 3–14, 2010.

[42] L. Jiang, Y. Du, Y. Zhang, B. R. Childers, and J. Yang, "LLS: cooperative integration of wear-leveling and salvaging for PCM main memory," *DSN*, pp. 221–232, 2011.

[43] S. Swami and K. Mohanram, "Reliable Nonvolatile Memories: Techniques and Measures," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 34, no. 3, pp. 31–41, 2017.

[44] Y. Wang, Y. Han, et al., "PSI Conscious Write Scheduling: Architectural Support for Reliable Power Delivery in 3D Die-Stacked PCM," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 5, pp. 1613–1625, 2016.

[45] H.-S. Wong, L. Heng-Yuan, Y. Shimeng, C. Yu-Sheng, W. Yi, C. Pang-Shiu, L. Byoungil, F. Chen, and T. Ming-Jinn, "Metal Oxide RRAM," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, May 2012.

[46] H. Manem, J. Rajendran, and G. S. Rose, "Design Considerations for Multilevel CMOS/Nano Memristive Memory," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 8, no. 1, February 2012.

[47] S. Hamdioui, H. Aziza, and G. C. Sirakoulis, "Memristor based memories: Technology, design and test," in *IEEE International Conference On Design and Technology of Integrated Systems In Nanoscale Era (DTIS)*, 2014.

[48] C. Ching-Yi, S. Hsiu-Chuan, W. Cheng-Wen, L. Chih-He, C. Pi-Feng, S. Shyh-Shyuan, and F. Chen, "RRAM Defect Modeling and Failure Analysis Based on March Test and a Novel Squeeze-Search Scheme," *IEEE Transactions on Computers*, vol. 64, no. 1, pp. 180–190, January 2014.

[49] N. Z. Haron and S. Hamdioui, "On defect oriented testing for hybrid CMOS/memristor memory," *Proceedings of the Asian Test Symposium*, pp. 353–358, 2011.

[50] S. Hamdioui, M. Taouil, and N. Haron, "Testing Open Defects in Memristor-Based Memories," *IEEE Transactions on Computers*, vol. 64, no. 1, pp. 247–259, October 2013.

[51] Y.-X. Chen and J.-F. Li, "Fault modeling and testing of 1T1R memristor memories," *VLSI Test Symposium*, 2015.

[52] N. Z. Haron and S. Hamdioui, "DfT schemes for resistive open defects in RRAMs," in *Design, Automation and Test in Europe Conference(DATE)*, 2012.

[53] A. Fantini, L. Goux, R. Degraeve, D. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y.-Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO2 RRAM," in *IEEE International Memory Workshop (IMW)*, 2013.

[54] P. Pouyan, E. Amat, and A. Rubio, "Statistical lifetime analysis of memristive crossbar matrix," in *IEEE International Conference On Design and Technology of Integrated Systems In Nanoscale Era (DTIS)*, 2015.

[55] C. An and L. Ming-Ren, "Variability of resistive switching memories and its impact on crossbar array performance," in *IEEE International Reliability Physics Symposium (IRPS)*, 2011.

[56] B. Chen, Y. Lu, B. Gao, Y. H. Fu, F. F. Zhang, P. Huang, Y. S. Chen, L. F. Liu, X. Y. Liu, J. F. Kang, Y. Y. Wang, Z. Fang, H. Y. Yu, X. Li, X. P. Wang, N. Singh, G. Q. Lo, and D. L. Kwong, "Physical mechanisms of endurance degradation in TMO-RRAM," in *IEEE International Electron Devices Meeting (IEDM)*, 2011.

[57] Y. Lu, B. Chen, B. Gao, Z. Fang, Y. Fu, J. Yang, L. Liu, X. Liu, H. Yu, and J. Kang, "Improvement of endurance degradation for oxide based resistive switching memory devices correlated with oxygen vacancy accumulation effect," in *IEEE International Reliability Physics Symposium (IRPS)*, 2012.

[58] Y. Y. Chen, B. Govoreanu, L. Goux, R. Degraeve, A. Fantini, G. S. Kar, G. Groeseneken, J. A. Kittl, D. J. Wouters, M. Jurczak, and L. Altimime, "Balancing SET/RESET Pulse for 10e10 Endurance in HfO2/Hf 1T1R Bipolar RRAM," in *IEEE International Electron Devices Meeting (IEDM)*, 2012.

[59] Y. Y. Chen, R. Degraeve, S. Clima, B. Govoreanu, L. Goux, A. Fantini, G. S. Kar, G. Pourtois, G. Groeseneken, D. J. Wouters, and M. Jurczak, "Understanding of the Endurance Failure in Scaled HfO2-based 1T1R RRAM through Vacancy Mobility Degradation," in *IEEE International Electron Devices Meeting (IEDM)*, 2012.

[60] D. Veksler, G. Bersuker, L. Vandelli, A. Padovani, L. Larcher, A. Muraviev, B. Chakrabarti, E. Vogel, D. Gilmer, and P. Kirsch, "Random telegraph noise (RTN) in scaled RRAM devices," in *IEEE International Reliability Physics Symposium (IRPS)*, 2013.

[61] P. Pouyan, E. Amat, S. Hamdioui, and A. Rubio, "RRAM Variability and Its Mitigation Schemes," in *International Workshop on CMOS Variability (VARI)*, 2016.

[62] P. Pouyan, E. Amat, and A. Rubio, "Memristive Crossbar Memory Lifetime Evaluation and Reconfiguration Strategies," *IEEE Transactions on Emerging Topics in Computing*, vol. PP, no. 99, pp. 1–1, June 2016.

[63] Y. Chen, H. H. Li, I. Bayram, and E. Eken, "Recent Technology Advances of Emerging Memories," *IEEE Design and Test*, vol. 34, no. 3, pp. 8–22, June 2017.

[64] S. Yu and P.-Y. Chen, "Emerging Memory Technologies: Recent Trends and Prospects," *IEEE Solid States Circuits*, vol. 8, no. 2, pp. 43–56, June 2016.

[65] 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 *IEEE International Electron Devices Meeting (IEDM)*, 2013.