Test of Supply Noise for Emerging Non-Volatile Memory

Mohammad Nasim Imtiaz Khan
School of Electrical Engineering and Computer Science
The Pennsylvania State University
University Park, U.S.A.
Email: muk392@psu.edu

Swaroop Ghosh
School of Electrical Engineering and Computer Science
The Pennsylvania State University
University Park, U.S.A.
Email: szg212@psu.edu

***Abstract*—Emerging Non-Volatile Memories (NVMs) suffer from high read/write current which can result in supply noise such as voltage droop and ground bounce. The magnitude of supply noise depends on the old data and the new data that is being written (for a write operation) or the stored data (for a read operation). In prior work, it has been shown that the noise generated by one access can affect another parallel access. Therefore, parallel read/write operation should be tested considering the supply noise. However, testing for read/write failure with supply noise considerations can take significant test time. In this work, we show that test time can be reduced by 410.82X for RRAM-based NVM Last Level Cache (LLC) by using Design for Test (DFT) circuits such as wordline overdrive and ending write operation early. We also show that the proposed test can save 79.875J of energy compared to the baseline test method.**

Keywords- Non-Volatile Memory, Supply Noise, Test, Test Time Reduction, Wordline Overdrive

I. INTRODUCTION

At the end of silicon roadmap, keeping the leakage power in tolerable limit has become one of the biggest challenges. Several promising Non-Volatile Memories (NVMs) are being investigated by the scientific community to address this issue. Emerging NVM technologies e.g., Spin-Transfer Torque RAM (STTRAM), Magnetic RAM (MRAM), Resistive RAM (RRAM), Phase Change Memory (PCM) and Ferroelectric RAM (FRAM) have drawn significant attention due to low (static) power operation, high density/speed and the inherent non-volatility [1-5]. Some of them have already entered the mainstream computing. Examples include MRAM by Everspin [6], CBRAM (a variant of RRAM) by Adesto Tech [7], PCM by Intel [8] and FRAM by Cypress [9]. However, their unique characteristics introduce new test challenges and call for designing new test methods and/or repurposing existing SRAM and Dynamic RAM (DRAM)/Embedded DRAM (eDRAM) test flows [10-15]. A well-defined test methodology will facilitate a broad adoption of these promising memory technologies in a variety of systems and applications.

Fig. 1(a) presents test flow for conventional memories.

This work is supported by Semiconductor Research Corp. (2847.001), NSF (CNS-1722557, CNS-1814710, CCF-1718474, DGE-1723687, and DGE-1821766) and DARPA Young Faculty Award (D15AP00089).

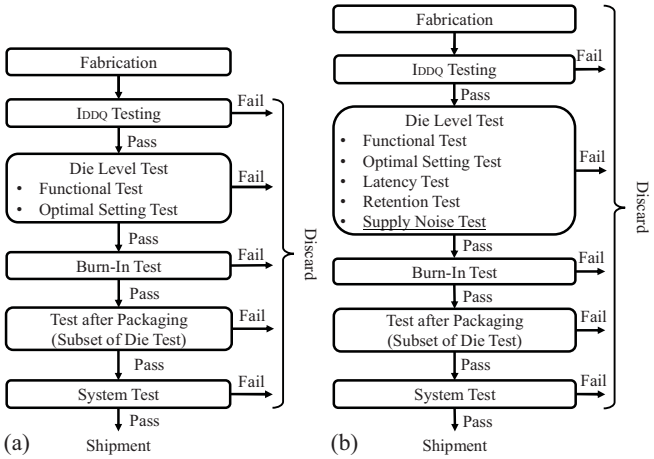

Fig. 1 Test flow: (a) conventional memory; (b) proposed for NVMs.

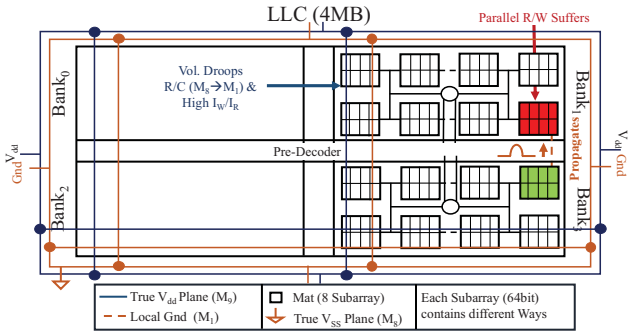

Fig. 2 1T1R-based 4MB LLC containing 4 banks showing supply noise. Each bank contains 8 Mats and each Mat contains 8 subarrays each producing 64bits. Each subarray has 8 Ways. Parallel read/write in $Bank_1$ (red) suffers due to propagation of droop/bounce from $Bank_3$ (green) (or vice versa).

Various March tests are performed to characterize read/write operations, address decoding and identify failures e.g., the coupling between neighboring cells. Tests to search for optimal setting identify optimal values of read/write assist techniques e.g., word-line over/under drive, supply voltage collapse [16] and negative bit-line [17]. For DRAM/eDRAM data retention is also characterized [18]. However, a similar retention test technique increases NVM test time significantly as NVMs can have a retention time of several years. In [10-12], weak-write-

978-1-5386-8382-8/18/$31.00 ©2018 IEEE

based retention test time compression technique is proposed for NVMs. In [13-14], external-magnetic-field-based retention test time compression technique is proposed for spintronics. In [19], a sneak-path-based testing is proposed to detect different faults such as stuck at faults or coupling faults. However, supply noise test for NVMs has not been presented before. The issue is described below:

Supply Noise [21]: Fig. 2 shows an overview of 4MB 1T1R-based Last Level Cache (LLC). Extremely high current (50-100mA assuming ~100µA/bit) is drawn from the supply for a full cache line (512-1024bit) write. This creates two issues [20-23]:

- *Supply voltage droop:* On-chip voltage regulator or power supply keeps the supply voltage constant. However, the supply voltage (distributed in higher metal layers such as M_9) reaches the memory bitcell (implemented in metal-1, M_1) via power-grid RC network. The interconnect resistance causes a significant voltage droop at the bitcell due to high current [20]. Voltage droop results in lower headroom for the bitcell and increases the write latency or decreases the sense margin for read. It can eventually lead to read/write failure [20-23].

- *Local ground bounce:* Similar to supply voltage, the true ground is routed on upper metal layer (e.g. M_8) and connects to the transistors in M_1. Therefore, the local ground rail bounces when the charge (due to high write/read current) is dumped.

Interestingly, the magnitude of combined supply noise (due to both droop and bounce) depends on the present state of the memory bit as well as the new data being written since I_{write} for 0→0, 0→1, 1→0 and 1→1 is different (for write operation), and on the stored data (for read operation) [20-23]. Read/write operation can be affected due to both self-inflicted and parallel read/write-inflicted combined supply noise. Therefore, bits should be tested and optimal V_{dd}/T_{clock} should be selected for successful read/write. However, traditional test approach fails to validate memory functionality for all possible corner cases (details in Section III).

The long write and read latency of emerging NVMs worsen the supply noise issue due to bank-level parallelism (i.e., read/write on independent banks in parallel) that is employed in LLC to achieve high bandwidth. Parallel access in emerging draw more current that can worsen the supply noise resulting in read/write failures. In summary, read/ write failure due to combined supply noise should be tested. However, exhaustive test for all possible read/write polarities and noise can degrade the test time significantly. Therefore, DFT circuits and test methodologies are required.

In this work, we describe the above NVM test challenges and propose new test methods and test patterns with associated Design-for-Test (DFT) circuits to solve these challenges. Fig. 1(b) presents a representative test flow for NVMs. Proposed new test method for NVMs is underlined in Fig. 1(b). For the sake of brevity, we restrict the discussion to one flavor of NVM namely, RRAM. *To best of our knowledge, this is the first attempt to study NVM-specific supply noise test challenges and corresponding solutions*.

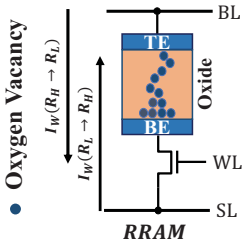

Fig. 3 Schematic of RRAM bitcell.

TABLE I: PARAMETERS USED FOR THE RRAM SIMULATION

Parameter	Value
Access Transistor W/L/V_T	130nm/65nm/0.423V
RRAM Oxide Gap for R_L/R_H	0.53nm/1.368nm
Cell Size	$12F^2$
Clock Frequency/V_{dd}	2GHz/2.2V
Read/Write Latency	0.5ns(1cycle)/10ns(20cycle)

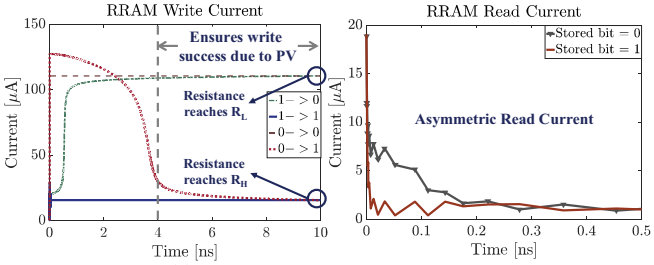

Fig. 4 RRAM high (a) write current; and (b) read current.

We make following contributions in this paper. We

(a) present new test challenges introduced by NVMs;

(b) summarize supply noise modeling from [21] due to NVM read/write operation and its impact on the parallel accesses;

(c) propose DFT techniques to reduce test time for the supply noise.

The paper is organized as follows: Section II presents the basics of NVMs. Sections III describes the background of supply noise for NVM. Sections IV describes test challenges associated with NVM under supply noise and propose techniques to address them. Section V presents discussion and Section VI draws conclusions.

II. BASICS OF NVM

In this section, we present the basics of NVM. We restrict our discussion to RRAM for the sake of brevity.

A. Basics of RRAM

RRAM contains an oxide material between Top/Bottom Electrode (TE/BE) (Fig. 3). RRAM resistive switching is due to oxide breakdown and re-oxidation which modifies a Conduction Filament (CF). Conduction through the CF is primarily due to transportation of electrons in the oxygen vacancies. These vacancies are created under the influence of

electric field due to the applied voltage. The two states of the RRAM are termed as Low Resistance State (LRS) and High Resistance State (HRS) and denoted by R_L/R_H respectively. The process of switching the state to LRS (HRS) is known as SET (RESET). We have used ASU RRAM Verilog-A model (bipolar HfO_x based resistive switching memory) [24] along with 65nm nMOS as an access transistor for simulation and analysis. All the model parameters are shown in Table I.

B. High Read/Write Current and Long Write Latency

RRAM suffers from long write latency (Fig. 4(a), ~10ns) since the current needed to switch the state is high (~100µA/bit). Read current for RRAM (~5.54µA/bit) (Fig. 4(b)) is also high compared to conventional memories.

C. Asymmetric Read/Write Current and Short Read Latency

Total read/write current for a full cache line is a function of data pattern due to asymmetric read/write current [25]. Therefore, the generated supply noise depends on the write/read data pattern. RRAM read latency can be optimized to less than 1ns (Fig. 4(b), ~0.5ns). Therefore, many reads can be initiated between one write to increase throughput (details in Section III).

III. Background of Supply Noise

A. Modeling of Voltage Droop/Ground Bounce [21]

Fig. 2 shows 4MB 1T1R LLC organization. It is a 4-way set associated cache. All the Ways of each Mat are accessed simultaneously and buffered at the edge of each Mat, resulting in a total of 512bit accesses. The figure also shows the upper layer metal plan. V_{dd} plane is in M_9 and V_{SS} plane is in M_8. Both V_{dd} and V_{SS} are implemented from M_7 to M_1 where M_7, M_5, M_3, and M_1 are horizontal and M_6, M_4, M_2 are vertical. The total area of the chip is 4970λ × 3950λ where each bank occupies 2046λ ×1536λ, and the remaining is occupied by the peripheral circuitry (e.g. pre-decoder, sense amp etc.). Note that λ is the feature size.

Ground Bounce: Fig. 5 shows the circuit used for ground bounce modeling. The total read/write current is dumped to the local ground implemented in M_1. Ground bounce propagates to nearest banks though metal M_1 via metal M_2, and then down to M_1 again. We modeled the resistance of path M_1 to M_8 by R_1. Fig. 6 shows the connection of true ground (M_8) with the local ground (M_1) of a Mat. We modeled the equivalent resistance

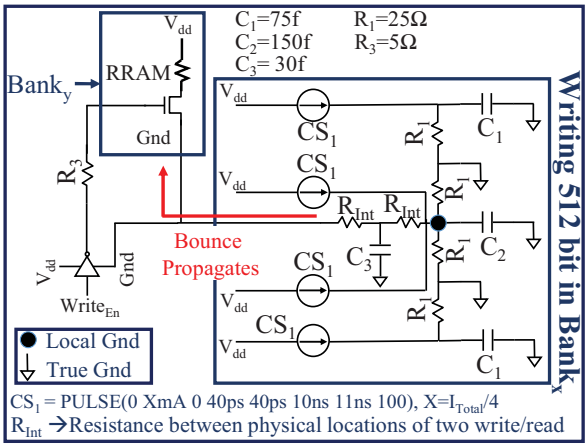

Fig. 5 Equivalent circuit for modeling ground (V_{SS}) bounce [21].

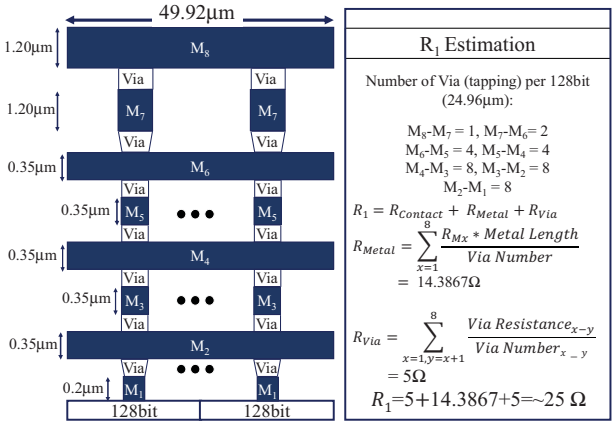

Fig. 6 Estimation of R_1 of Fig. 5 for ground bounce modeling [21].

using 65nm layout parameters (Table II) [26-27]. We divided 512bits into 4 groups (only two of them shown in Fig. 6) for simplicity. Each metal layer R/C and via between metal layers are also given in Table II. Our estimation shows that R_1 is equivalent to ~25Ω (Fig. 6). Fig. 7(a) shows that as R_1 increases, ground noise increases. Capacitance calculation is omitted for the sake of brevity.

The impact of write current (per bit) and the width of write data on ground noise are shown in Fig. 7(b) and Fig. 7(c) respectively. Fig. 7(b) shows that as write current (per bit) increases, ground noise increases. Typically, NVM write

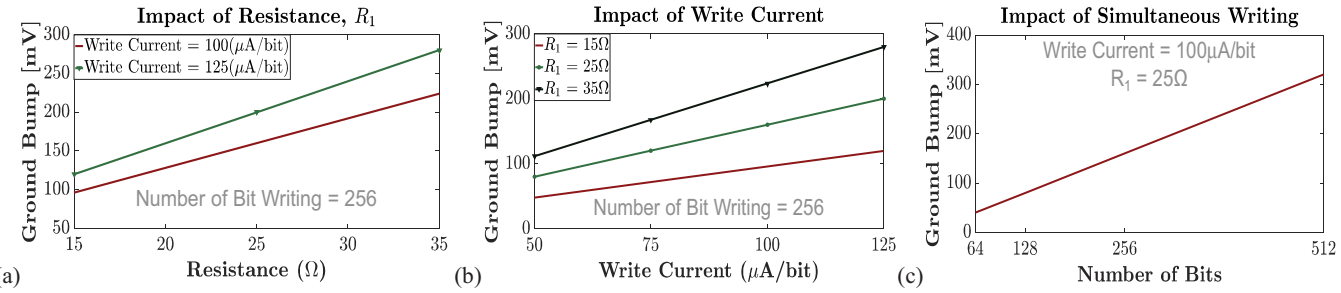

Fig. 7: (a) Impact of resitance (R_1) of Fig. 5 on the local ground bounce [23]; (b) impact of write current (per bit) on the local ground bounce [23]; and, (c) impact of number of bits writing on the local ground bounce [23].

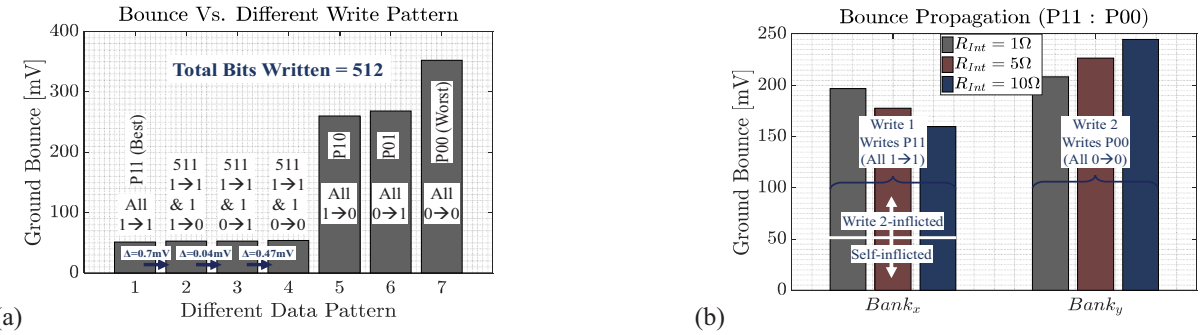

Fig. 8 (a) Ground bounce vs write data pattern [21]; and, (b) impact of R_{Int} on bounce when Write-1/Write-2 writes P11/P00 pattern in $Bank_x$/$Bank_y$ respectively [21]. Write-1 observes higher bounce as R_{Int} reduces, even though Write-1 generates only ~51.42mV of self-bounce.

TABLE II: PARAMETERS USED FOR GROUND BOUNCE MODELING

Parameter	Value
Resistance (Ω/μm) $M_1/M_2/M_3/M_4/M_5/M_6/M_7/M_8$	0.91/0.41/0.41/0.41/0.41 /0.41/0.04/0.04 [26-27]
Cap (fF/μm) $M_1/M_2/M_3/M_4/M_5/M_6/M_7/M_8$	0.13/0.17/0.17/0.17/0.17 /0.17/0.19/0.19 [26-27]
Miller Coupling Factor (MCF)	1.5
Via Resistance (Ω) $M_{1\text{-}2}/M_{2\text{-}3}/M_{3\text{-}4}/M_{4\text{-}5}/M_{5\text{-}6}/M_{6\text{-}7}/M_{7\text{-}8}$	6/5/5/3/3/1/1 (CVD Tungsten-based) [28]
Di-electric Constant for Cap. Calculation (C_{plate}/C_{side})	2.2/2.79 [27]
Resistance between M_1 to Source/Drain Contact, $R_{Contact}$ (Ω)	~5 [29]

current varies from 50μA to 125μA. Fig. 7(c) shows that the ground noise increases as the width of write data in a memory array increases.

We also model the resistance R_{int} which represents the equivalent resistance between the local ground of one address of a bank to another address of another bank. Our estimation shows that lowest (closest two addresses of two banks)/highest (furthest two addresses of two banks) R_{int} is 1.63Ω/185.12Ω. Average read/write current for a full cache line is divided into four constant Current Sources (CS). Therefore, current magnitude of CS, XmA is equal to $I_{Total}/4$ (for example, 512bit of RRAM 0→0 writing, I_{Total}=56.32mA and X=14.08mA) and each one presents read/write current for 128bit.

Fig. 8(a) shows the bounce generated by a full cache line write of RRAM employed in this work for various data patterns. It is notable that 1→1 (we call it P11) write creates lowest (~51.42mV) (best-case), and 0→0 (we call it P00) creates highest (~352.46mV) (worst-case) ground bounce. It is also evident that other data patterns create a bounce in between the bounce generate by P11 and P00. Furthermore, the bounce generated can be at the granularity of 1mV for various data patterns. Fig. 8(b) shows the bounce observed by Write-1 when Write-1 writes P11 in $Bank_x$ and Write-2 writes P00 in $Bank_y$ at the same time. The bounce equalizes each other as R_{Int} reduces. For example, Write-1 observes ~196.72mV (for R_{Int} =1Ω) even though Write-1 generated only ~51.42mV of self-bounce by writing P11 pattern.

Voltage Droop: High write current creates droop due to the presence of interconnect resistance between the power supply source and destination. This is especially true for the farthest Mats of the cache as it incurs highest parasitic resistance (i.e. highest droop). Simulation indicates that supply voltage can droop to ~0.9V_{dd} when writing P00 to all the bits to $Bank_3$ (Fig. 9(a)). Droop is modeled using a circuit model similar to Fig. 5 (details omitted for brevity).

B. Parallel Read/Write Operation [21]

NVM write latency requires multiple clock cycle. For example, the required number of clock cycles is 5 (1) with a clock frequency of 2GHz and write (read) latency of 2.5ns (0.5ns). However, the throughput will degrade if the memory access is completely stopped during 5 or 1 cycles (Fig. 9(b)-(c)). In practice, RRAM write latency is even higher (10ns, ~20 cycle for this work). Therefore, parallelism is used to perform write/read simultaneously to different banks and increase

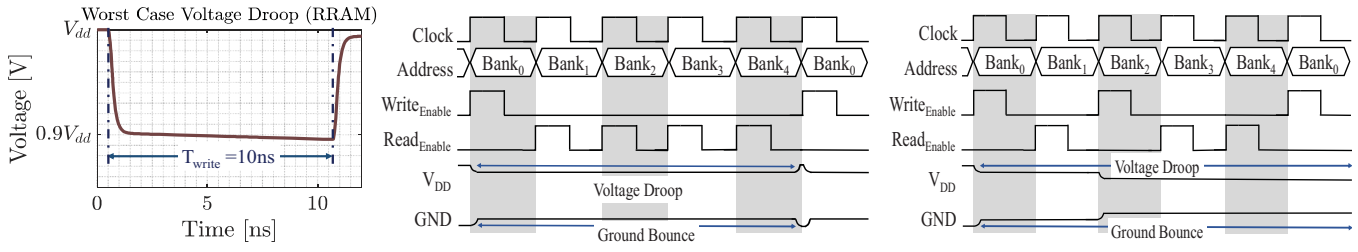

Fig. 9 (a) Worst-case (P00) voltage droop when writing all the bits to one MAT of $Bank_3$ (Fig. 2) [21]; (b) four reads are initiated between two writes. We call it 1X mode [21]; (c) three reads and one write are initiated between two writes. We call it 2X mode [21].

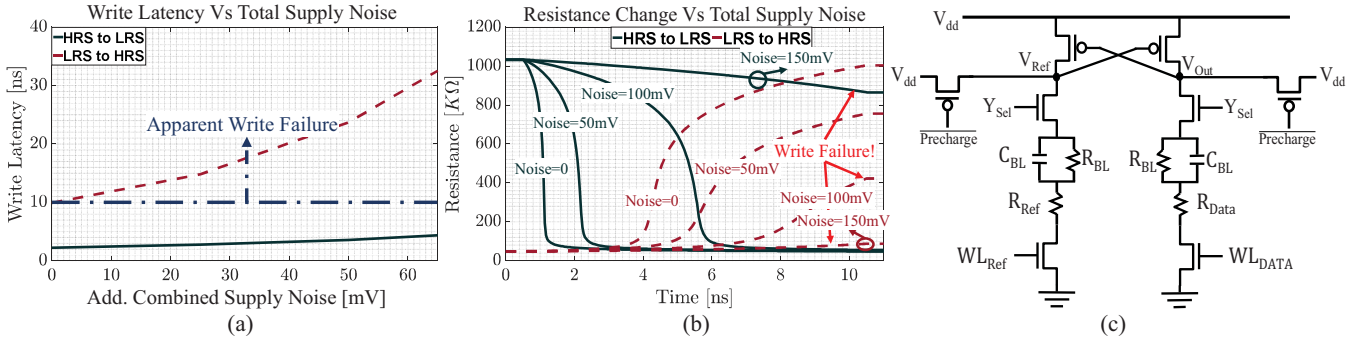

Fig. 10 (a) RRAM write latency increases as additional supply noise increases [21]; (b) RRAM resistance variation with additional supply noise [21]; and, (c) single ended read circuitry used it this work [30]. We considered R_{BL} = 25Ω, R_{Ref} = 500KΩ and C_{BL} = 25fF [21].

system throughput. Parallel access can take following forms:

1X Write: Read can be initiated in the next 4 cycles in other banks (Fig. 9(b)) when one write has been initiated in one bank. These data are processed in the pipeline to maintain high throughput. Read operations initiated in cycles 2, 3, 4 and 5 will experience failure (due to supply noise propagation to those banks) owing to, (a) poor sense margin at a lower voltage; (b) higher access transistor resistance at lower wordline voltage. We call this write/read scheme 1X write.

nX Write: Multiple (n) writes can be initiated with read. For example, one write along with 3 consecutive reads can be initiated in the next four clock cycles in other banks (Fig. 9(c)) when write has been initiated in one bank. The second write will draw additional current from the supply which might add to the existing noise. Furthermore, the local ground will bounce due to the second write along with multiple reads and propagate to the first write (or vice versa) location and cause write failure. We call this write/read scheme 2X write.

In this work, we have ignored additional droop (due to insignificant magnitude) and considered only the ground bounce caused by a read operation. However, both ground bounce and droop are considered as the noise components generated by a write operation.

C. Parallel Write Failure Due to Supply Noise [21]

We simulate RRAM write operation with additional supply noise (excluding self-inflicted noise). It is evident from Fig. 10(a) that as supply noise increases, write latency for both LRS to HRS and HRS to LRS increases. However, the former increases very rapidly compared to the later. At this point, we can consider that even with 10mV of voltage loss, LRS to HRS write fails as the corresponding write latency is around 12ns (>10ns). However, for better understanding, let's consider Fig. 10(b) which shows the RRAM resistance changing during write operation with respect to supply noise. Note that HRS to LRS write operation can sustain up to 100mV (accurately ~120mV) supply noise. On the contrary, the final resistance of LRS to HRS does not reach the full R_H (=1000KΩ) value for even 50mV (reaches till 760KΩ). However, we can still consider this as successful write since sufficient sense margin will be generated during read operation of this bit (using read circuitry of Fig. 10(c)). This is true since the final resistance is greater than R_{Ref} (=500KΩ). Additional voltage loss beyond 50mV and less than 120mV will cause LRS to HRS write failure but still can write HRS to LRS successfully. Therefore, write operation is:

i) successful, if the write incurs an additional supply noise < 50mV from parallel read/write operation;

ii) partially successful i.e., only HRS to LRS (1→0) is successful if the write operation incurs an additional supply noise > 50mV but < 120mV from parallel read/write operation;

iii) unsuccessful, if the write operation incurs an additional supply noise > 120mV from parallel read/write operation.

Now, we analyze write failure due to supply noise from

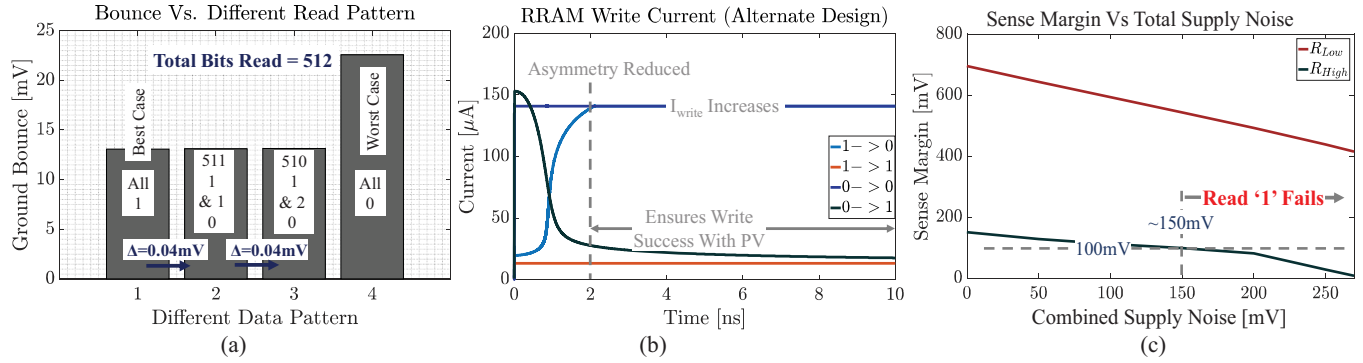

Fig. 11 (a) Bounce generation vs different read data pattern [21]; (b) RRAM current profile for alternate design (symmetric) [21]; and, (c) sense margin with additional supply noise [21]. Sense margin for data 1 suffers more, and failure is observed above 150mV of additional supply noise.

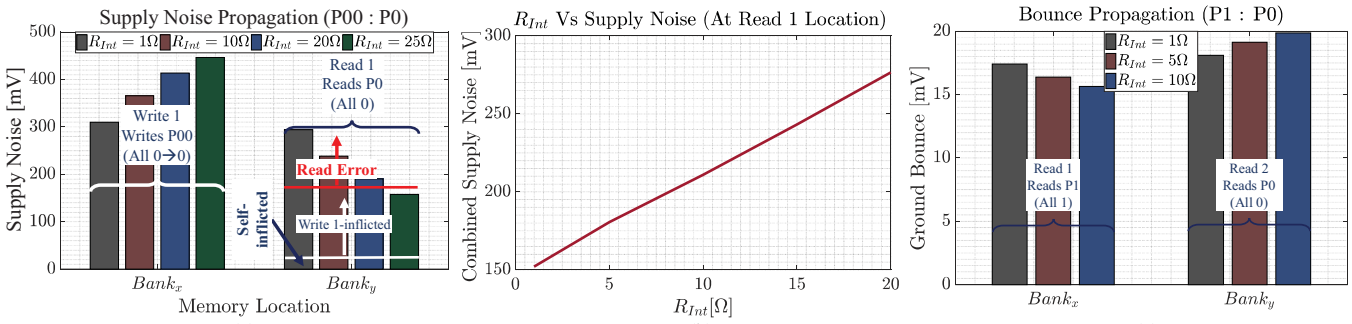

Fig. 12 (a) Write-1/Read-1 are performed with P00/P0 data pattern in $Bank_x$/$Bank_y$ respectively [21]. For this particular case, Read-1 can be affected by Write-1 if R_{int} between the parallel access is < 23Ω; (b) R_{Int} vs supply noise generated by Write-1 that causes Read-1 failure [21]; and, (c) Read-1/Read-2 are performed in $Bank_x$ /$Bank_y$ with P1/P0 read pattern respectively [21]. Parallel reads does not affect each other.

parallel read/write operation:

Write Failure Due to Parallel Write: Result indicates that if Write-1/Write-2 writes P00/P11 pattern respectively, both polarity write operation fails if R_{int} between them is $0 < R_{Int} < 22\Omega$. Furthermore, only one type of polarity write operation fails (0→1) if R_{int} between them is $22\Omega < R_{Int} < 37\Omega$. It should be noted that for this example, Write-1 and Write-2 are generating the lowest and the highest self-inflicted supply noise respectively. However, if Write-1 generates more self-supply noise, the write operation failure happens for larger R_{int}.

Write Failure Due to Parallel Read: Fig. 11(a) shows the ground bounce generation by read operation as a function of read data pattern. The figure shows that the maximum bounce generated by the read operation is ~23mV. Therefore, it is evident that write failure by the parallel read is not possible even if their physical location is right next to one another.

The RRAM employed in this work takes longer latency for writing 0→1. Therefore, only 0→1 write operation fails if it incurs a certain range of supply noise ($22\Omega < R_{Int} < 37\Omega$). However, we further analyzed another type of RRAM design which eliminates the asymmetric write current by using asymmetric doped transistor [31]. Fig. 11(b) shows the current profile for all four cases for such alternate designs. I_{write} for 1→0 and 0→0 increases (more supply noise) although asymmetry is almost eliminated (by reversing TE/BE of RRAM cell, reducing access transistor V_T by 100mV and using separate write voltages for $V_{1\to0}$=2V and $V_{0\to1}$=1.8V). We still kept the write time 10ns to successfully write even with process variation. We performed a similar investigation on this circuit for write failure. We observe that (for this symmetric design) if the write operation incurs 132mV of noise with a period of 1ns, and the noise continues for the entire write operation (10ns), only 1→0 write is successful. Furthermore, if the write operation incurs 116mV of noise with a period of 10ns, only 0→1 write is successful. Therefore, symmetric designs could be worse as both polarity write failure is possible based on different supply noise condition. Note that 1ns-wide supply noise can be generated by a read operation (along with write for higher magnitude) while 10ns-wide supply noise can be generated by a write operation.

D. Parallel Read Failure Due to Supply Noise [21]

Read circuitry: Fig. 10(c) presents the single-ended read circuity [30] used in this work. V_{ref} and V_{Out} are both kept at V_{dd} by pre-charging through p-MOS M_3/M_4, and thereby p-MOS M_1, M_2 are both OFF. R_{BL} (= 25Ω) and C_{BL} (= 25fF) are bitline resistance and capacitance respectively. Y_{sel}, $WL_{Ref,}$ and WL_{DATA} are enabled during a read operation. If R_{Data} is greater than R_{Ref} (=500KΩ), V_{Out} remains at V_{dd} while V_{Ref} discharges to zero. However, if R_{Data} is less than R_{Ref} (= 500KΩ), V_{Out} discharges to zero while V_{Ref} remains at V_{dd}.

We analyzed the sense margin generated by the read circuitry shown in Fig. 10(c) for both data 0 and 1. Fig. 11(c) shows that the sense margin reduces with the increment of supply noise. However, ground bounce affects more compared to V_{dd} droop as it, i) reduces the discharge current, and ii) reduces V_{GS} of access transistor (resulting increment of $R_{Transistor}$) while voltage droop only reduces discharge current. Note that if the read operation in a bank incurs additional supply noise > 150mV from a parallel read/write in another bank, the operation reads 1 incorrectly (Fig. 11(c). However, sense margin for data 0 is above 150mV even with 350mV of supply noise. Therefore, read 0 does not fail even for the worst possible additional supply noise caused by another parallel read/write.

Read Failure Due to Parallel Write: We analyzed read failure in one bank due to parallel write operation in another bank. Fig. 12(a) shows that Write-1 is performed in $Bank_x$ with P00 write pattern and Read-1 is performed in $Bank_y$ with P0 read pattern (P0 means all the bits stored in that address contain data 0). Results indicate that read fails in $Bank_y$ by Write-1 in $Bank_x$ if R_{Int} between Write-1 and Read-1 is < 23Ω. This observation is true if Write-1 generates worst-case supply nosie (caused by P00 write pattern). However, if Write-1 is of different data pattern which generates lower supply noise compared to the worst-case, R_{int} range for possible read failure will be lowered. This means that Write-1 in $Bank_x$ affects Read-1 in $Bank_y$ for a lesser memory area. Fig. 12(b) shows R_{Int} vs supply noise generated by Write-1 that causes Read-1 failure.

Read Failure Due to Parallel Read: We analyzed read failure in one bank due to parallel read operation in another bank. Fig. 12(c) shows that Read-1 is performed in $Bank_x$ and Read-2 is performed in $Bank_y$ with P1/P0 read pattern

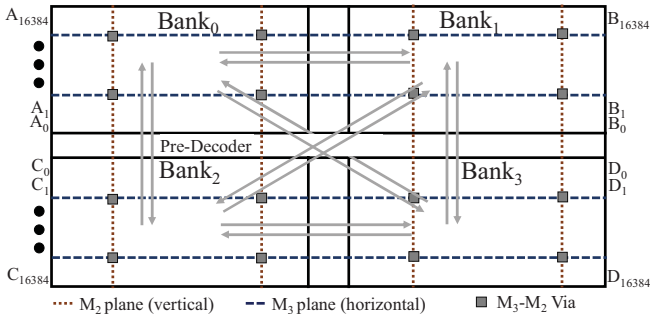

Fig. 13 4MB LLC diagram showing addresses of $Bank_0$/$Bank_1$/ $Bank_2$/$Bank_3$ as A/B/C/D from 0 to 16K. The grey arrows show different cases for testing the impact of supply noise.

respectively. It can be concluded that parallel read operations in $Bank_x$ and $Bank_y$ does not affect each other for any value of R_{int}.

IV. NVM Test with Supply Noise

In this section, we present the challenges of NVM test in the presence of high supply noise and propose techniques to test supply noise-induced read/write failure.

A. Test Data Patterns & Test Cases

Data Pattern for Write Test: The worst-case for supply noise-induced write/read failure depends on the physical implementation of the memory. The data pattern (especially for write) is also another factor that determines the supply noise. Therefore, we need to select data pattern in a way that generates maximum supply noise while detecting failures. For example, writing 0→0 generates the highest supply noise. However, write failure cannot be detected as old data and new data are the same. Therefore, we propose to write data pattern in a way that only one bit of the address incurs 0→1 writing and rest of the bits incurs 0→0 writing. For example, test data pattern can be 0x00000000→ 0x00000001 for a 32-bit cache line for testing write failure. Note that the test pattern mentioned above only tests the LSB bit of that address. The test should be repeated 31 times with appropriate test data pattern to test the rest of the 31-bits of the same address.

Data Pattern for Read Test: We need to detect read failure for the stored-data = 1. If all the bits of an address are 1, all bits can be tested at once. However, the supply noise will be reduced by 19mV from the worst-case (when all bits of the address are 0) for a 512bit cache line. Therefore, we can test a few 1's in the data when the rest of the bits are 0's. For example, the read-test-data pattern can be 0x00110011 for a 32-bit cache line. This data pattern tests 4 bits of the address. The same address needs to be tested 7 more times to test the rest of the 28bits with appropriate data patterns (stored data in the bits of interest = 1). However, write failure occurs for only 50mV of supply noise whereas read failure occurs for 150mV. Therefore, reads are robust against supply noise compared to write and we can test only the write failures to minimize the test time.

Fig. 13 shows the physical location of the addresses of different banks $Bank_0$/$Bank_1$/$Bank_2$/$Bank_3$ named as A/B/C/D respectively from 0 to 16K. Note that this banking architecture is for illustrative purpose. We assume that only two concurrent writes and many reads are possible in independent banks. Therefore, all the banks can be accessed in parallel with 2 reads and 2 writes or with 4 reads. Grey arrow of Fig. 13 depicts various cases for testing the impact of supply noise. We summarize all the possible test cases below:

Case-1 (Accesses in adjacent banks): This case includes the impact of writing in one bank (e.g. $Bank_0$) and writing to the adjacent bank (e.g. $Bank_1$). Another adjacent bank pair is $Bank_2$-$Bank_3$. Write failure should be tested by writing to the address pairs (A_0-B_0, A_1-B_1, …, A_{16384}-B_{16384}, C_0-D_0, C_1-D_1, …, C_{16384}-D_{16384}) simultaneously and detecting write failure in both addresses. Note that the abovementioned test can be done on diagonal addresses on these bank pair (for e.g. A_0-B_1, A_0-B_2, …, A_0-B_{16384}, A_1-B_0, A_1-B_2, …, A_1-B_{16384}, … etc.). However, diagonal addresses will experience less noise due to a higher R_{int} and will not be the worst-case. Therefore, we can skip their test and focus on the worst-case addresses.

Case-1 may seem like the worst-case supply noise. However, the bounce generated at metal layer M_1 needs to go to M_2, then M_3, then propagate through M_3 and going down to M_1 via M_2. Therefore, R_{int}, in this case, is higher for the same physical distance.

Case-2 (Accesses in physically confronting banks): This case includes the impact of writing in one bank (e.g. $Bank_0$) and writing to the bank on top or bottom (e.g. $Bank_2$). Another confronting bank pair is $Bank_1$-$Bank_3$. Write failure should be tested by writing to the address pairs (A_0-C_0, A_0-C_1, …, A_0-C_X, A_1-C_0, A_1-C_1, …, A_1-C_{X-1}, …, A_{x-1}-C_0 and similarly for $Bank_1$-$Bank_3$) simultaneously and detecting write failure in the both addresses. It should be noted that A_X-C_X/B_X-D_X represents the furthest addresses that can affect each other.

Case-3 (Accesses in diagonal banks): This case captures the impact of writing in one bank (e.g. $Bank_0$) and writing to the diagonal bank (e.g. $Bank_3$). Another diagonal bank pair is $Bank_1$-$Bank_2$. Write failure should be tested by writing to the address pairs (A_0-D_0, A_0-D_1, …, A_0-D_X, A_1-D_0, A_1-D_1, …, A_1-D_{X-1}, …, A_{x-1}-D_0 and similarly for $Bank_1$-$Bank_2$) simultaneously and detecting write failure in both addresses. Note that A_Y-B_Y/C_Y-D_Y represents the furthest addresses that can affect each other.

Case-3 incurs highest R_{int} (addresses are physically distant) and the bounce needs to propagate through metal layer M_3 like Case-1. We call two addresses of two banks corresponding to Case-1/Case-2/Case-3 as address-pairs as the impact of supply noise needs to be tested on them during parallel accesses to them.

C. Proposed Test Time Compression Technique & Algorithm

Testing all possible cases requires significant test time which is not acceptable for mass production. Therefore, techniques for test time compression are required. It should be noted that the main reason for high test time is the long write latency. Therefore, we propose the following techniques to

reduce the write latency during testing:

i) Wordline Overdrive (WL OV): This technique is implemented by increasing the wordline voltage during write in test mode. WL OV reduces the access transistor resistance. Therefore, the bits may draw more current with the same bitline and sourceline voltage. Higher I_{write} may change the supply noise profile and test mode might not match with the operation mode. Fig. 14(a) shows RRAM I-V curve with 500mV of WL OV for both 0→1 and 1→0 compared to the case of Fig. 4(a). Write latency reduces (4ns from 10ns) and write current increases for both writes.

We mentioned earlier that we need 0→1 transition to test a bit and we need to test each bit of all addresses. This means that a single address needs to be tested 512 times (for 512bit cache line) with appropriate test pattern. We call this number N_{repeat}. WL OV can help to reduce this number. Average write current for 1→0, 0→0 and 0→1 increases to 118.16µA, 126.72µA and 90.57µA from 80.08µA, 110.01µA and 84.34µA respectively. Therefore, test pattern with combination of the 0→0, 0→1 and 1→0 write patterns can be selected to maintain the same worst-case write current as 512-writes of 0→0 as normal mode (without WL OV). This approach reduces N_{repeat}. A rough estimation shows that each 0→1 reduces total I_{write} (from worst-case) by 26µA whereas one 0→0 and one 0→1 increases total I_{write} by (8+16.7)µA=24.7µA. This means that 170 writes of 0→1, 172 writes of 0→0 and 170 writes of 1→0 (in a 512bit data) yield almost the same total I_{write} as 512-writes of 0→0. Therefore, one address can be tested 3 times (N_{repeat} = 3) where

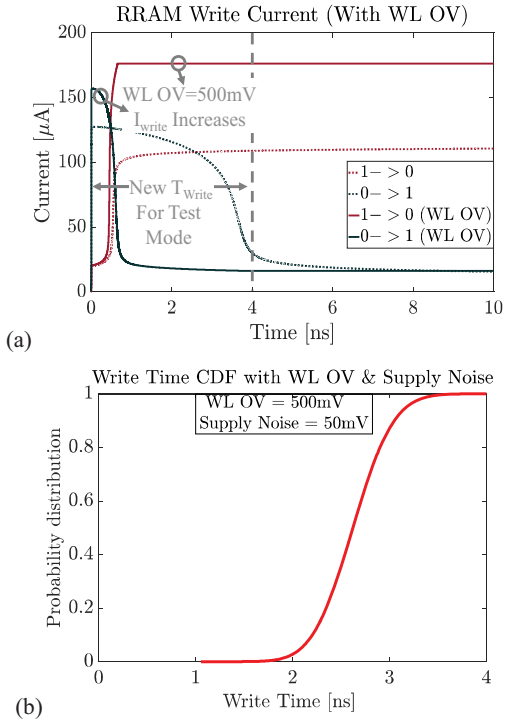

Fig. 14 (a) RRAM I-V curve with 500mV of WL OV for both 0→1 and 1→0. Write latency reduces and write current increases for both writes; and, (b) write time CDF with 500mV of WL OV and 50mV of additional supply noise. All the bits pass with write time = 4ns.

170, 171 and 171 bits can be tested each time.

ii) Ending write operation early: We observe that the main part of the write operation is completed in first 4ns/2ns for the RRAM shown in Fig. 4(a)/Fig. 11(b) respectively. The write time is still kept as 10ns for two reasons: i) the resistance value reaches the target HIGH state (for writing 0→1); and, ii) weakest bit (due to process variation) gets written. We have mentioned earlier that the resistance of the cell is not required to reach the target HIGH state (Fig. 10(b)) during 0→1 writing. If the resistance crosses the resistance of the reference cell by a margin (we considered the margin as 75KΩ), the bit can still be read correctly within the target read latency (0.5ns in our case). Furthermore, write time variation can be tightened by WL OV. Therefore, the memory bits can be written successfully with a shorter time by WL OV and by ending the write operation early.

We implemented 500mV of WL OV and 50mV of additional supply noise and did a 300-point Monte-Carlo analysis on write operation with 3σ of 5% of oxide gap for R_L with a mean of gap = 0.530n. The result shows that all bits pass write operation successfully with 4ns of write time (Fig. 14(b)). If the supply noise is increased, the write time needs to be increased to achieve the desired final resistance (75KΩ of margin over reference resistance) for a successful read operation. Therefore, WL OV during test mode can reduce the write time to 4ns and still model the supply noise generated during normal operation mode.

It can be noted that architectural choices can also reduce supply noise impact and reduce test time significantly. For example, parallel accesses can be allowed in a way that Case-2 never needs to be tested. This can be achieved by disabling parallel accesses in $Bank_0$-$Bank_2$, $Bank_1$-$Bank_3$ pairs.

Proposed test algorithm: Let's call the required old data as $Init_1$/$Init_2$/$Init_3$ and new data as TP_1/TP_2/TP_3 to test 170/171/171 bits respectively in a 512bit cache line. For example, for an 8bit cache line one possible choice can be $Init_1$= 0b00111**000,** $Init_2$= 0b00**000**111, $Init_3$=0b**00**111111 and TP_1=0b00000**111,** TP_2=0b00**111**000 and TP_3=0b**11**000000. This will test 3bits/3bits/2bits respectively while Init→TP writing in all cases are almost similar to a worst-case total write current (8bits of 0→0 in normal operation mode). Note that the bold digits indicate the bits of interest during each test. The steps (each one is performed sequentially) to test write failure due to supply noise from parallel writes are:

1. Write $Init_1$ to one address of an address-pair (let's say A_0);
2. Write $Init_1$ to the other address of address-pair (let's say B_0);
3. Write $Init_1$ to one address of second address-pair from another bank-pair (let's say C_0);
4. Write $Init_1$ to the other address of second address-pair from another bank-pair (let's say D_0);
5. Read A_0/B_0/C_0/D_0 simultaneously to verify write success of $Init_1$ (write success of Step 1, 2, 3 and 4);
6. Write TP_1 to 1st address-pair simultaneously (A_0-B_0);

7. Write TP_1 to 2nd address-pair simultaneously (C_0-D_0);
8. Read $A_0/B_0/C_0/D_0$ simultaneously to verify write success of TP_1 (write success of Step 6 and 7);
9. Repeat steps 1 to 8 with $Init_2$-TP_2 and then, $Init_3$-TP_3 patterns for A_0-B_0 and C_0-D_0 address-pairs;
10. Repeat steps 1 to 9 for all address-pairs of Case-1;
11. Repeat step 10 for Case-2 and then Case-3.

Step-6/7 tests both addresses A_0 and B_0 / C_0 and D_0 at the same time for supply noise-induced write failure. Step 5/8 minimizes the test time by sharing read for all the banks. Supply Noise Test Algorithm shows Case-1 testing which can also be used for Case-2/Case-3 by replacing N with X/Y respectively. The symbol (**:**) means the operations are performed in parallel. The proposed algorithm is given below:

Supply Noise Test Algorithm
Case-1
for(i=0, i<=N, i++)
{ for(j=1, i<=3, j++)
{ $(wInit_j)_{Ai}$ *;* $(wInit_j)_{Bi}$ *;* $(wInit_j)_{Ci}$ *;* $(wInit_j)_{Di}$ *;*
$(rZ_1)_{Ai} : (rZ_2)_{Bi} : (rZ_3)_{Ci} : (rZ_4)_{Di}$ *;*
If $(Z_1, Z_2, Z_3, Z_4 \neq Init_j)$ *exit();*
$(wTP_j)_{Ai} : (wTP_j)_{Bi}$ *;*
$(wTP_j)_{Ci} : (wTP_j)_{Di}$ *;*
$(rZ_1)_{Ai} : (rZ_2)_{Bi} : (rZ_3)_{Ci} : (rZ_4)_{Di}$ *;*
If $(Z_1, Z_2, Z_3, Z_4 \neq TP_1)$ *exit();*
}
}

D. Test Time Analysis

The value of X and Y for testing Case-2 and 3 depends on the memory implementation. We estimate X = 4916 and Y = 4096 for our implementation. Total number of addresses in a bank is N (= 16K in our case). The number of Bank-pairs in all cases is N_{pair} (= 2 in our case) and N_{repeat} is 3 (512 without the proposed compression technique).

The total test time for testing supply noise impact can be formulated as below:

$$T_{Total} = T_{Case-1} + T_{Case-2} + T_{Case-3}$$

$$T_{Case-1} = N_{repeat} * 2 * N_{pair} * N * T_{write} + N_{repeat} * N_{pair} * N * T_{write} + 2 * N * N_{repeat} * T_{read}$$

The first term of T_{Case-1} is for writing Initial data pattern to both addresses of address-pair (Step1, 2, 3, 4), the second term is for parallel writing with test pattern (Step 6, 7) and the third term is for write success validation (Step 5,8). We have ignored the read time for the last address-pair (= 0.5ns) and read time for checking the initial data pattern write success (= $N*N_{repeat}*0.5$ns). T_{Case-2} and T_{Case-3} can be formulated similarly by replacing N with $\frac{X(X+1)}{2}$ and $\frac{Y(Y+1)}{2}$. We estimate that T_{Case-1}, T_{Case-2} and T_{Case-3} are 0.511s, 383.657s and 262.057s in base-case (without compression) which totals to 646.225s. Typically, total target test time for a chip is 2-3sec [32]. Therefore, the base-case test time is not acceptable. However, T_{Case-1}, T_{Case-2} and T_{Case-3} reduces to 1.229ms, 0.9427s and 0.6293s which totals to 1.573s if write time is reduced using the proposed technique. Therefore, 410.82X test time compression can be achieved with the proposed technique.

Test energy: In the proposed test method, we incur almost the same total write current with same bitline/sourceline voltage (different wordline voltage is not an issue as gate current increment is insignificant) and incurs lower test time due to lower write time. Therefore, the proposed method saves a total energy of

$$V_{dd} * I_{write} * \Delta Test_{Time_{compression}}$$

$$= 2.2\text{V} * 56.32\text{mA} * 644.652\text{s}$$

$$= 79.875J$$

V. Discussion

A. Design Techniques to Mitigate Supply Noise [21]:

Following design techniques can prevent or alleviate the supply noise impact on parallel read/write operation:

i) Sequential read/write access: A naïve solution is to implement non-pipelined access only. However, this hurts the system throughput as several clock cycles will be required to execute one read/write operation.

ii) Intelligent architecture: Parallel–operations of different processes can be initiated to addresses with highest possible R_{Int}. However, this will alleviate the issue to some extent only.

iii) Good quality V_{dd}/V_{ss} grid: A good V_{dd}/V_{ss} grid reduces R_1 (in Fig. 5) which in turn reduces supply noise. However, as the technology is scaled down, R_1 might increase eventually. Therefore, this technique cannot eliminate the issue.

iv) Power rail separation for each bank: Separation of supply and gnd rails between parallel accessed banks will prevent propagation of supply noise. However, this will incur significant area/design overheard (power regulators and separate metal grids needed). Furthermore, separating the rail reduces the rail capacitance which is not desirable (high power rail capacitance is desired for supply noise cancellation).

v) Slowing system clock: Higher T_{Clock} gives more time to read/write at lower headroom voltage to fix latency failures. However, T_{clock} must be at least twice (2X throughput loss) to prevent write failure for just 80mV of noise (result extended from Fig. 10(b)).

Above techniques can alleviate supply noise. However, the issue might still persist as NVMs incur high process variation. Therefore, there might be some weak bits which will still be

vulnerable to parallel-access-inflicted supply noise.

B. Consideration to other NVMs:

Although the study on supply noise impact is carried out for RRAM LLC, we believe that a broad range of NVMs is going to face this issue. Therefore, NVMs should be tested for supply noise-induced error. Note that the proposed test technique can be implemented for other NVMs.

VI. Conclusion

In this work, we summarize that high write current of NVM can lead to supply noise which propagates to the neighboring banks and can affect parallel read/write. Therefore, NVMs should be tested for supply noise-inflicted errors. We propose test techniques to maximize and catch supply noise induced failures. Our analysis indicate that test time and energy could prohibitively high to validate all possible test cases. Therefore, we propose test compression techniques such as, wordline overdrive and early write termination to reduce supply noise test time. We also suggested design techniques to minimize supply noise.

References

[1] A. Nigam, C. W. Smullen, V. Mohan, E. Chen, S. Gurumurthi, and M. R. Stan, "Delivering on the promise of universal memory for spin-transfer torque ram (STT-RAM)," in IEEE/ACM International Symposium on Low Power Electronics and Design, pp. 121–126, Aug 2011.

[2] D. C. Worledge, G. Hu, P. L. Trouilloud, D. W. Abraham, S. Brown, M. C. Gaidis, J. Nowak, E. J. O'Sullivan, R. P. Robertazzi, J. Z. Sun, and W. J. Gallagher, "Switching distributions and write reliability of perpendicular spin torque MRAM," in 2010 International Electron Devices Meeting, pp. 12.5.1–12.5.4, Dec 2010.

[3] Y. Wu, S. Yu, X. Guan, and H. S. P. Wong, "Recent progress of resistive switching random access memory (RRAM)," in 2012 IEEE Silicon Nanoelectronics Workshop (SNW), pp. 1–4, June 2012.

[4] A. Pirovano, A. L. Lacaita, F. Pellizzer, S. A. Kostylev, A. Benvenuti, and R. Bez, "Low-field amorphous state resistance and threshold voltage drift in chalcogenide materials," IEEE Transactions on Electron Devices, vol. 51, pp. 714–719, May 2004.

[5] Y. M. Kang and S. Y. Lee, "The challenges and directions for the mass production of highly-reliable, high-density 1T1C FRAM," in 2008 17th IEEE International Symposium on the Applications of Ferroelectrics, vol. 1, pp. 1–2, Feb 2008.

[6] "16Mb 256K x 16 MRAM Memory - Everspin." https://www.everspin.com/file/882/download, 2015. [Online; accessed May-03-2018].

[7] "RM24C256DS, 256-Kbit 1.65V Minimum Non-volatile Serial EEPROM I2C Bus. http://www.adestotech.com/wpcontent/uploads/RM24C256DS085.pdf, 2016. [Online; accessed May-03-2018].

[8] "Intel Optane Memory Series." https://ark.intel.com/products/97544/Intel-Optane-Memory-Series-16GB-M_2-80mm-PCIe-3_0-20nm-3D-Xpoint [Online; accessed May - 03 - 2018].

[9] "FM28V102A 1-Mbit (64 K x 16) F-RAM Memory." http://www.cypress.com/file/140901/download. [Online; accessed May-03-2018]

[10] A. Iyengar, S. Ghosh, and S. Srinivasan, "Retention testing methodology for sttram," IEEE Design Test, vol. 33, pp. 7–15, Oct 2016.

[11] I. Yoon, A. Chintaluri, and A. Raychowdhury, "Emacs: Efficient MBIST architecture for test and characterization of stt-mram arrays," 2016 IEEE International Test Conference (ITC), pp. 1–10, 2016.

[12] H. Naeimi, C. Augustine, A. Raychowdhury, S.-L. Lu, and J. Tschanz, "STTRAM scaling and retention failure," vol. 17, pp. 54–75, 01 2013.

[13] M. N. I. Khan, A. S. Iyengar, and S. Ghosh, "Novel magnetic burn-in for retention and magnetic tolerance testing of STTRAM," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 26, pp. 1508–1517, Aug 2018.

[14] M. N. I. Khan, A. S. Iyengar and S. Ghosh, "Novel magnetic burn-in for retention testing of STTRAM," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, Lausanne, 2017, pp. 666-669.

[15] M. N. I. Khan and S. Ghosh, "Test challenges and solutions for emerging non-volatile memories," in 2018 IEEE 36th VLSI Test Symposium (VTS), pp. 1–6, April 2018.

[16] K. Kim, H. Jeong, J. Park, and S. Jung, "Transient cell supply voltage collapse write assist using charge redistribution," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 63, pp. 964–968, Oct 2016.

[17] S. Mukhopadhyay, R. M. Rao, J. Kim, and C. Chuang, "SRAM writeability improvement with transient negative bit-line voltage," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 19, pp. 24–32, Jan 2011.

[18] H. Yang, C. Chang, M. C. . Chao, R. Huang, and S. Lin, "Testing methodology of embedded DRAMs," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 20, pp. 1715–1728, Sept 2012.

[19] S. Kannan, J. Rajendran, R. Karri, and O. Sinanoglu, "Sneak-path testing of crossbar-based nonvolatile random access memories," IEEE Transactions on Nanotechnology, vol. 12, pp. 413–426, May 2013.

[20] R. K. Aluru and S. Ghosh, "Droop mitigating last level cache architecture for STTRAM," in Proceedings of the Conference on Design, Automation & Test in Europe, DATE '17, (3001 Leuven, Belgium, Belgium), pp. 262–265, European Design and Automation Association, 2017.

[21] M. N. I. Khan and S. Ghosh, "Fault injection attacks on emerging non-volatile memory and countermeasures". In Proceedings of the 7th International Workshop on Hardware and Architectural Support for Security and Privacy (HASP '18). ACM, New York, NY, USA, 2018.

[22] M. N. I. Khan and S. Ghosh. 2018. Information Leakage Attacks on Emerging Non-Volatile Memory and Countermeasures. In Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED '18). ACM, New York, NY, USA, Article 25, 6 pages.

[23] M. N. Khan, and S. Ghosh, "Analysis of Row Hammer Attack on STTRAM," 2018 IEEE 36th International Conference on Computer Design (ICCD), Orlando, FL, USA, 2018.

[24] P. Y. Chen and S. Yu, "Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design," IEEE Transactions on Electron Devices, vol. 62, pp. 4022–4028, Dec 2015.

[25] K. Shamsi and Y. Jin, "Security of emerging non-volatile memories: Attacks and defenses," in 2016 IEEE 34th VLSI Test Symposium (VTS), pp. 1–4, April 2016.

[26] "Interconnect: Capacitance and Resistance for 65nm technology." http://ptm.asu.edu/, 2005. [Online; accessed May-03-2018].

[27] "Wire Capacitance and Resistance Calculator for 65nm." http://users.ece.utexas.edu/~mcdermot/vlsi-2/Wire_Capacitance_and_Resistance_65nm.xls, 2008. [Online; accessed May-03-2018].

[28] I. Shao, J. M. Cotte, B. Haran, A. W. Topol, E. E. Simonyi, C. Cabral, and H. Deligianni, "An alternative low resistance MOL technology with electroplated rhodium as contact plugs for 32nm CMOS and beyond," in 2007 IEEE International Interconnect Technology Conference, pp. 102–104, June 2007.

[29] X. Li, W. Zhao, Y. Cao, Z. Zhu, J. Song, D. Bang, C. C. Wang, S. H. Kang, J. Wang, M. Nowak, and N. Yu, "Pathfinding for 22nm CMOS designs using Predictive Technology Models," in 2009 IEEE Custom Integrated Circuits Conference, pp. 227–230, Sept 2009.

[30] T. D. Happ, H.-I. Lung, N. T. Nirschl, "Current Compliant Sensing Architecture For Multilevel Phase Change Memory", US 11/620,432, 2007.

[31] S. H. Choday, S. K. Gupta and K. Roy, "Write-Optimized STT-MRAM Bit-Cells Using Asymmetrically Doped Transistors," in IEEE Electron Device Letters, vol. 35, no. 11, pp. 1100-1102, Nov. 2014.

[32] "Test time." https://semiengineering.com/addressing-test-time-challenges/. [Online; accessed May-03-2018].