Fault Injection Attacks on Emerging Non-Volatile Memory and Countermeasures

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

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

ABSTRACT

Emerging Non-Volatile Memories (NVMs) suffer from high and asymmetric read/write current and long write latency which can result in supply noise such as supply voltage droop and ground bounce. The magnitude of supply noise depends on the old data and the new data that is being written (for write operation) or on the stored data (for read operation). In this paper, we show that the adversary can write specific data pattern (that results in deterministic supply noise) in their memory space to launch, i) Denial of Service (DoS) attack (total write failure), and ii) specific polarity fault (i.e., fault injection) attack in victim's memory space sharing the same power rails with the adversary's memory space. These attacks are specifically possible if exhaustive testing of the memory for all patterns, all possible location combinations, all possible parallel read/write conditions are not performed under bit-to-bit process variations and, specified (-10°C to 90°C) and unspecified temperature ranges (i.e., less than -10°C and greater than 90°C). Simulation result indicates that adversary can launch DoS attack on victim's write operation by injecting more than 120mV of supply noise to victim's write location. The adversary can also launch $0 \rightarrow 1$ polarity fault injection attack on victim's write operation by injecting supply noise greater than 50mV but shorter than 120mV to victim's write location. Furthermore, the adversary can cause data '1' read failure by injecting more than 150mV of supply noise to victim's read location.

CCS CONCEPTS

• **Security and privacy → Hardware attacks and countermeasures**; • **Hardware → Memory and dense storage**;

KEYWORDS

Emerging NVM, Security, Privacy, Fault Injection, DoS

ACM Reference Format:
Mohammad Nasim Imtiaz Khan and Swaroop Ghosh. 2018. Fault Injection Attacks on Emerging Non-Volatile Memory and Countermeasures. In *Proceedings of Hardware and Architectural Support for Security and Privacy (HASP'18).* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3214292.3214302

HASP '18, June 2, 2018, Los Angeles, CA, USA

ACM ISBN 978-1-4503-6500-0/18/06…$15.00
https://doi.org/10.1145/3214292.3214302

1 INTRODUCTION

At the end of Silicon roadmap, keeping the leakage power in tolerable limit has become one of the biggest challenges. Several emerging Non-Volatile Memories (NVMs) are being investigated by the scientific community to address this issue. Emerging NVM technologies for example, 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, high 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) [7] by Adesto Tech, PCM by Intel [8] and FRAM by Cypress [9].

Application of Emerging NVMs: STTRAM can reach the speed and the endurance of SRAM and therefore can replace SRAM in L2/L3 cache [10, 11]. Both STTRAM and RRAM are proposed to replace eFlash [10, 12]. PCM-based Solid State Drive (SSD), namely Optane is already sold by Intel [8]. Furthermore, NVM enables low-power computation and novel architecture [10, 11, 13]. NVM has been also investigated for application beyond memory and proposed for novel applications such as neuromorphic computing, ambient sensor, security primitive etc. [10, 13]. Although these memories are promising, their unique characteristics introduce new threats to data security and data privacy which are not present in conventional memories like SRAM or DRAM. Therefore, it is necessary to investigate their vulnerabilities before their mass adaption.

Emerging NVM vulnerabilities: Most of the emerging NVMs are susceptible to ambient parameters such as, temperature and magnetic field which can be used to launch Denial-of-Service (DoS) attacks [14, 15]. In [16], it has been pointed out that NVMs suffer from asymmetric and high read/write current (i.e., read/write current for data '0' and data '1' are different) which can be exploited to launch Side-Channel Attack (SCA) [17]. NVMs also suffer from supply voltage droop due to high read and write current [18]. However, the ground bounce phenomena and the role of supply noise (i.e. droop and ground bounce) on security and privacy have not been explored before.

In this paper, we present that high and asymmetric write current and long write latency of emerging NVMs cause supply noise such as supply voltage droop and ground bounce. We further show that an adversary can leverage the supply noise to launch fault injection attacks. The adversary can write specific data pattern (i.e., specific number of 0's and 1's in the write data) to generate deterministic supply noise. The noise will propagate to victim-user's memory space and cause failure of read/write operation. This fault injection can leak system assets such as, keys during sensitive operation such

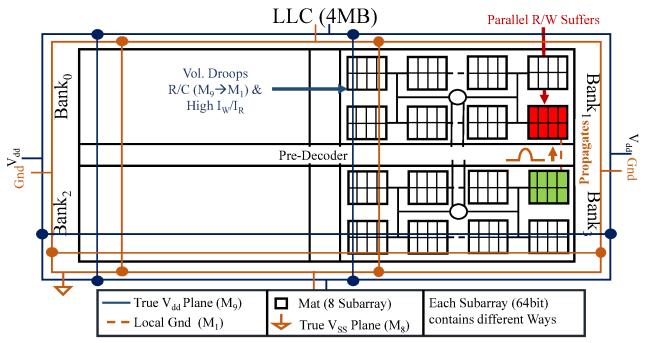
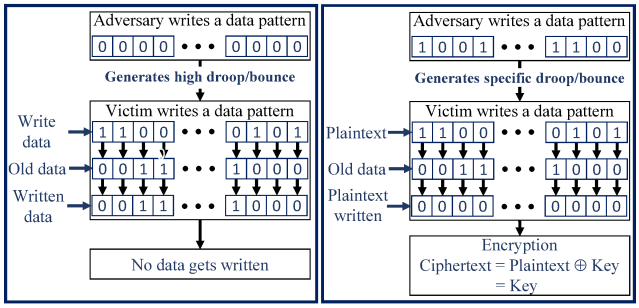


Figure 1: 1T1R-based 4MB LLC (containing 4 banks) showing supply noise (droop and bounce). Each bank contains 8 Mats and each Mat contains 8 subarrays each producing 64bits. Each subarray has 8 Ways. Parallel read/write in $Bank_1$ (red) suffers due to propagation of supply noise from $Bank_3$ (green)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



(a) DoS Attack (b) Fault Injection Attack

Figure 2: (a) DoS attack; (b) specific polarity fault injection attack.

as encryption. For example, a simple XOR encryption module takes a plaintext and XORs it with the key to generate the ciphertext. If an adversary can set all the bits of the plaintext to 0 (1) by injecting fault, the ciphertext becomes same as keys (1's complement of key). Therefore, the adversary can recover the key and figure out the plaintexts in the consecutive cycles (given that the key is not changed).

We have considered 1T1R RRAM-based Last Level Cache (LLC) in this work as a test case. However, the attack is also applicable to other NVMs such as, STTRAM as they consume high write/read current and incur long write latency. Therefore, the observations made in this paper are generic for NVM LLCs. It is notable that RRAM is usually considered for main memory owing to poor endurance. However, high endurance (~10^{12}) RRAM has been also proposed recently [19], which can be suitable for LLC.

Attack model and assumptions: In this work, we have assumed the followings:

(i) NVM LLC is being shared by two users and the users are an adversary and a victim;
(ii) Adversary has the knowledge of the amount of supply noise that can be generated by a read/write data pattern during read/write operation initiated by him;
(iii) Adversary knows the propagation model of the generated supply noise (decays with distance) and the impact of the propagated noise on the victim's read/write operation;
(iv) The adversary is an expert in computer architecture and can exploit knobs such as, accessing specific data pattern in a predefined physical locations to prevent their replacement by policies such as, Least Recently Used (LRU).

Fig. 1 shows an overview of 4MB 1T1R-based LLC. Extremely high current (~50mA assuming 100µA/bit) is drawn from the supply for a full cache line (512bit) write operation. This creates two issues:

- Supply voltage droop: On-chip voltage regulator or power supply keeps the supply voltage constant. However, the supply voltage, V_{dd} (distributed in metal M_9) reaches the memory bitcell (implemented in metal M_1) via power-grid RC network. The interconnect resistance causes a significant voltage droop at the bitcell due to high current drawn by the bitcell during read/write operation. Voltage droop causes lower headroom for the bitcell and increases the write latency for write operation or decreases the sense margin for read operation. It can eventually lead to read/write failure.
- Local ground bounce: The true ground (V_{SS}) is routed on upper metal layer (for example, M_8) and connects to the transistors in M_1 (similar to V_{dd} routing). Therefore, the voltage of local ground rail bounces when the charge (due to high write/read current) is dumped to it.

The magnitude of total supply noise (droop and bounce) depends on the present state of the memory bit as well as the new data being written since write current for $0 \rightarrow 0$, $0 \rightarrow 1$, $1 \rightarrow 0$ and $1 \rightarrow 1$ are different (for write operation), and on the stored data (for read operation). It should be noted that read/write operation can be affected due to both self-inflicted and parallel read/write-inflicted supply noise. Therefore, bits are tested, and optimal V_{dd}/T_{clock} are selected for successful read/write. However, traditional test approach may fail to validate memory functionality for all possible corner cases. The adversary can leverage this to add supply noise to victim's location, and affect victim's read/write (details provided in Section III).

The long write and read latency of emerging NVMs worsen the supply noise issue due to bank-level parallelism (i.e., parallel reads/writes on independent banks) that is employed in LLC to achieve high bandwidth. Parallel accesses in emerging NVM-based LLC draw more current which can worsen the noise issue resulting in read/write failure. Therefore, an adversary can exploit this vulnerability to launch fault inject attacks on parallel accesses. Furthermore, the nature of cache (for example, set associativity) and the replacement policies (for example, LRU) can force adversary's address space physically close to the victim's address space making the task easier for the adversary. A detailed approach to manipulate addresses using cache associativity and replacement policy is beyond the scope of this paper (subject of our future research). In summary, adversary can launch following attacks by leveraging the noise:

DoS attack: Fig. 2a shows the concept for launching DoS attack. Adversary can store a data pattern which creates very high magnitude of supply noise (for example, $0 \rightarrow 0$) during write operation. This noise will propagate to victim's memory space and

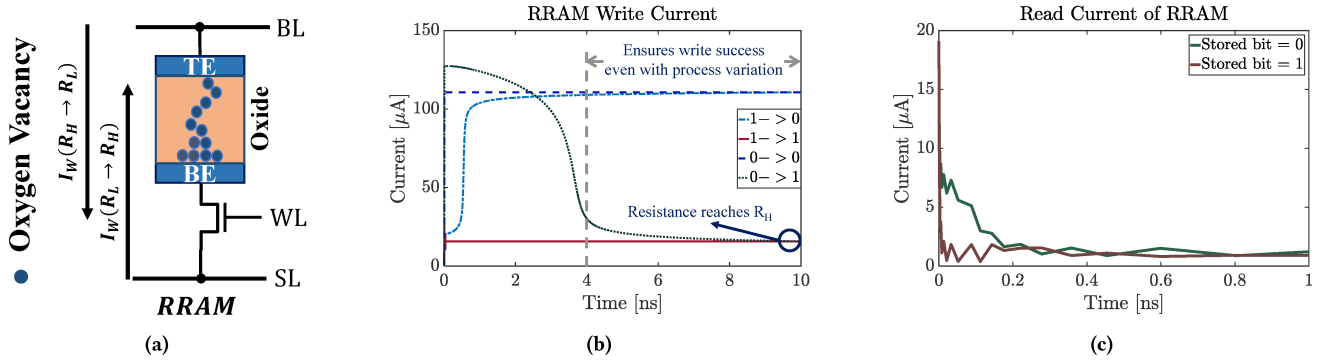


Figure 3: RRAM (a) bitcell; high (a) write current; and (b) read current.

cause complete read/write failure to result in system failure (i.e., DoS).

Fault injection attack: Fig. 2b shows the concept for launching specific polarity fault injection. Adversary can write a specific data pattern (i.e. specific magnitude of noise) to cause one specific polarity write failure. For example, 10.4mA of write current is needed to generate 130mV of supply noise. Now, the adversary can store 0x0000FFFFFFFFFFFFFFFFFFFFFFFFFFFF pattern in his memory space, and write this pattern to an address (after flashing the address) to generate 130mV of noise (assuming $0 \rightarrow 1/0 \rightarrow 0$ write consumes 80μA/100μA respectively), and cause $0 \rightarrow 1$ write failure of a near-by parallel access. We observe that $1 \rightarrow 0$ can be written successfully whereas $0 \rightarrow 1$ write fails for RRAM if certain magnitude of supply noise is generated in other banks by a parallel write. We note that the adversary can control the polarity of write error in victim's memory space. This can be further utilized to strengthen SCA for key extraction. Fault injection attacks are well-known in the security community [20] and implementation of fault injection in Multi-Level Cell (MLC) NAND flash has been studied [21]. However, its implementation in emerging NVM has not been explored before.

Following contributions are made in this paper:

(a) We show that high write current of emerging NVM can lead to ground bounce which propagates to the neighboring banks;
(b) We show that supply noise worsen the write latency and sense margin if write/read is performed in parallel in other banks. Therefore, those bits can suffer from failures;
(c) We model the droop/bounce for RRAM during write/read, and show the magnitude of supply noise required to create specific polarity write failure and DoS attack;
(d) We show the magnitude of supply noise required to cause read failure;
(e) We propose potential design-level countermeasures.

Rest of the paper is organized as follows: Section II presents the background on RRAM, high and asymmetric read/write current, supply noise generation, and its impact on parallel read/write operation; Section III describe DoS and fault injection attacks; Section IV presents discussion on practicality of the proposed attacks and design level mitigation techniques; Section V draws conclusion.

2 BACKGROUND

In this section, we present the basics of RRAM. We also describe droop/bounce modelling, and their relation to fault injection.

2.1 Basics of RRAM

RRAM contains an oxide material between Top/Bottom Electrode (TE/BE) (Fig. 3a). 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 the transportation of electrons in the oxygen vacancies. These vacancies are created under the influence of electric field due to applied voltage. The two states of the RRAM are termed as Low Resistance State (LRS), R_L and High Resistance State (HRS), R_H. The process of switching the state to LRS (HRS) is known as SET (RESET). We have used ASU RRAM Verilog-A model [22] along with 65nm nMOS as access transistor for simulation and analysis. The RRAM is bipolar HfO_x-based resistive switching memory [22]. All the model parameters used in this work are shown in Table 1.

High read/write current and long write latency: RRAM suffers from long write latency (Fig. 3b, 10ns) and the write current

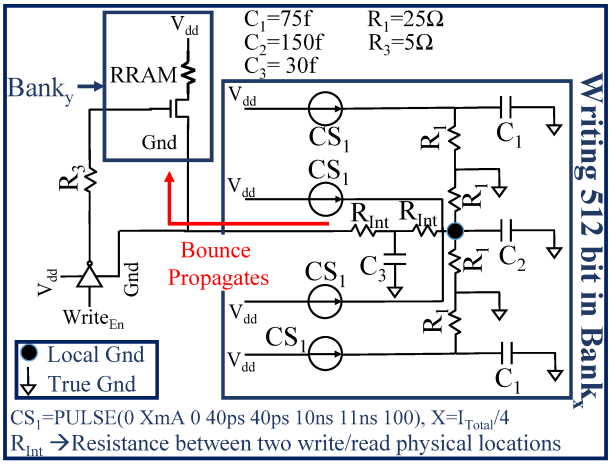


Figure 4: Equivalent circuit for modeling local ground bounce.

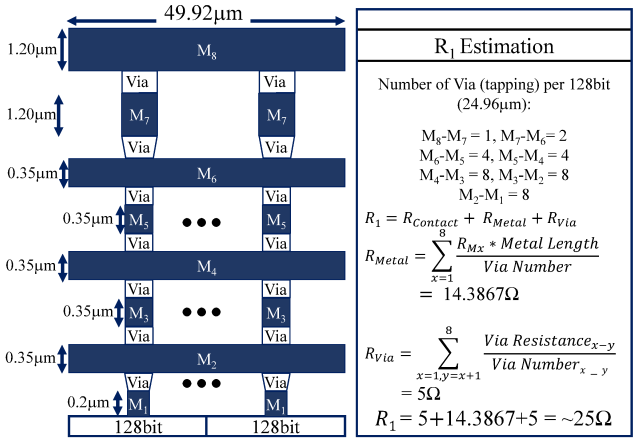


Figure 5: Estimation of R_1 (of Fig. 4) for ground bounce modeling.

required to switch the state is high (~100µA/bit). Read current for R-RAM (~5.54µA/bit) is also high compared to conventional memories (Fig. 3c).

Asymmetric read/write current: Due to asymmetric write current [16], the total write current for a full cache line is a function of data pattern. An adversary can select the write data pattern to precisely control and generate supply noise. Furthermore, NVM read current is also asymmetric (Fig. 3c).

Table 1: Parameters Used for the Simulation

Parameter	Value
Access Transistor W/L/V_T	195nm/65nm/0.423V
RRAM Gap for R_L/R_H	0.53nm/1.368nm
Unit Cell Size	$12F^2$
System Clock Frequency/V_{dd}	2GHz/2.2V
Read/Write Latency	0.5ns(1cycle)/10ns(20cycle)

Table 2: 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 [23, 24]
Capacitance (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 [23, 24]
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) [25]
Di-electric Constant for Cap. Calculation (C_{plate}/C_{side})	2.2/2.79 [24]
Res. between M_1 to Source/ Drain Contact, $R_{Contact}$ (Ω)	~5 [26]

2.2 Modeling of Voltage Droop & Ground Bounce

Fig. 1 shows 4MB 1T1R LLC organization. It is a 4-way set associated cache. All the Ways of each Mat are accessed simultaneously and

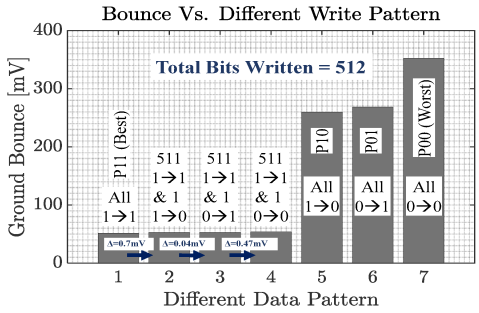


Figure 6: Local ground bounce vs write data pattern.

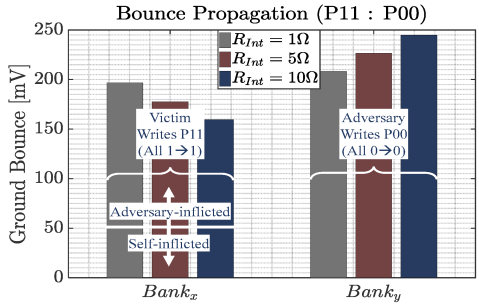


Figure 7: Impact of R_{Int} on ground bounce when victim/adversary writes P11/P00 in $Bank_x$/$Bank_y$ respectively. Victim observes higher bounce as R_{Int} reduces, even though victim generated only ~51.42mV of self-bounce.

buffered at the edge of each Mat, resulting in 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} is 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 $4970\lambda \times 3950\lambda$ where each bank occupies $2046\lambda \times 1536\lambda$ and the remaining is occupied by the peripheral circuitry (for example, pre-decoder etc.). Note that λ is the feature size.

Ground bounce: Fig. 4 shows the circuit used for ground bounce modeling. The total read/write current is dumped to the local ground implemented in M_1 and causes bounce of local ground voltage. Ground bounce propagates to nearest banks through 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. 5 shows the connection of true ground (M_8) with the local ground (M_1) of a address of a Mat. We modeled the equivalent resistance using 65nm layout parameters (Table II) [23, 24]. We divided 512bits to 4 groups (only two of them shown in Fig. 5) for simplicity. Each metal layer R/C and via resistance between metal layers are also given in Table II. Our estimation shows that R1 is equivalent to ~25Ω (Fig. 5). Magnitude of ground bounce depends on this value. Capacitance calculation is omitted for brevity. We also modeled the resistance R_{int} which represents the equivalent resistance between the local ground of one address of one bank to the local ground of another address of another bank. Our estimation shows that the lowest (closest two addresses of two banks) R_{int} is 1.63Ω. Average read/write current for a full cache line is divided into four constant Current Sources (CS). Therefore, current magnitude

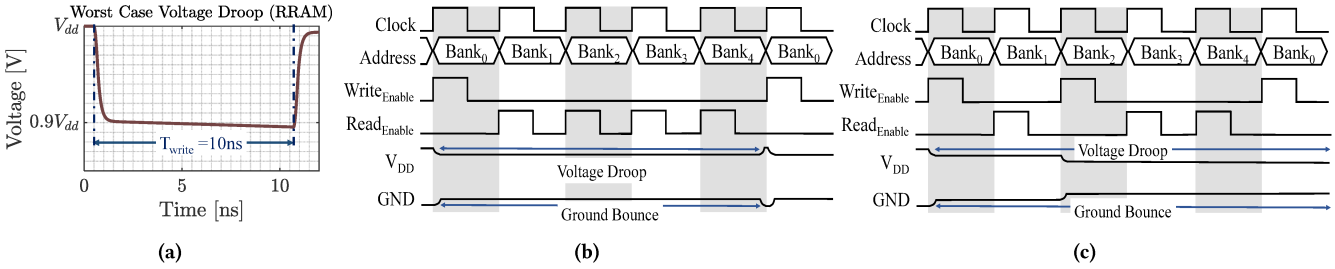


Figure 8: Worst-case (P00) voltage droop when writing all the bits to one Mat of $Bank_3$ (Fig. 1); (b) four reads are initiated between two writes. We call it 1X write; and, (c) three reads and one write are initiated between two writes. We call it 2X write.

of CS, XmA is equal to $I_{Total}/4$ (for example, 512bits of $0 \rightarrow 0$ writing, I_{Total}=56.32mA and X=14.08) and each one presents total read/write current for 128bits.

Fig. 6 shows the bounce generated by a full cache line write for various write data patterns. It is notable that $1 \rightarrow 1$ (we call it P11) write creates lowest (~51.42mV) (best-case), and $0 \rightarrow 0$ (we call it P00) creates highest (~352.46mV) (worst-case) ground bounce. It is also evident that other data patterns create bounce in between P11 and P00. Furthermore, the bounce can be controlled at the granularity of 1mV by choosing corresponding write data pattern. Fig. 7 shows the bounce observed by victim when victim write P11 in $Bank_x$ and adversary writes P00 in $Bank_y$ at the same time. The bounce equalizes each other as R_{Int} reduces, and victim observes ~196.72mV (for R_{Int} =1Ω) even though victim generated only 51.42mV of self-bounce by writing P11.

Voltage droop: High write current creates supply voltage droop due to the presence of interconnect resistance between the power supply source (implemented in M_9) and destination (bitcell, implemented in M_1). 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 of a Mat of $Bank_3$ (Fig. 8a). Droop is modeled using a circuit model similar to Fig. 4 (details omitted for brevity).

2.3 Parallel Read/Write Operation

RRAM write latency requires multiple clock cycles. For example, the required number of clock cycles are 5 (1) with a system clock frequency of 2GHz and write (read) latency of 2.5ns (0.5ns). However, the throughput will degrade if memory access is completely stopped during 5 or 1 cycles (Fig. 8b-8c). In practice, RRAM write latency is even higher (10ns, 20 cycle for this work). Therefore, parallelism is used to perform write/read in successive cycles (can be initiated by different users). The parallel access can take following forms:

1X write: Read can be initiated in the next 4 cycles in other banks (Fig. 8b) for adversary when a write has been initiated for victim. These data are processed in 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 lower voltage headroom; (b) higher access transistor resistance at lower word-line 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. 8c) for adversary when a write has been initiated for victim. The second write will draw additional current from the supply which might add to the existing supply noise. Furthermore, local ground will bounce due to second write along with multiple reads and propagate to first write (or vice versa) location and cause write failure. We call this write/read scheme 2X write.

2.4 Supply Noise

In the rest of the paper, we have combined the magnitude of droop and bounce and call it as supply noise. It should be noted that supply noise can be both, self-inflicted and parallel-read/write-inflicted. We call the parallel–read/write-inflicted supply noise as additional supply noise. We have ignored the droop caused by read operation (due to insignificant magnitude), and considered only the ground bounce. Therefore, supply noise generated by read operation is effectively ground bounce only. However, both droop and bounce component of supply noise are considered for write operation.

3 DoS & FAULT INJECTION ATTACK

In this section, we discuss DoS and fault injection attack methodology by leveraging supply noise.

3.1 DoS & Fault Injection by Write Failure

We have simulated RRAM write operation with additional supply noise (excluding self-inflected noise) injected by parallel read/write operation. It is evident from Fig. 9a that as the additional supply noise increases, write latency for both LRS to HRS ($0 \rightarrow 1$) and HRS to LRS ($1 \rightarrow 0$) increases. However, the write latency for the former one increases very rapidly compared to the later one. At this point, we can consider that LRS to HRS write fails with even 10mV of additional supply noise as the corresponding write latency is around 12ns (>10ns). However, let's consider Fig. 9b for better understanding which shows the RRAM resistance switching during write operation with respect to additional supply noise. It is evident that HRS to LRS write operation can sustain up to 100mV (accurately 120mV) of additional supply noise. On the contrary, final resistance of LRS to HRS does not reach the full R_H (=1000KΩ) (reaches till 760KΩ) value for even 50mV of supply noise. However, we can still consider this as successful write since sufficient sense margin will be generated during the read operation of this bit (using

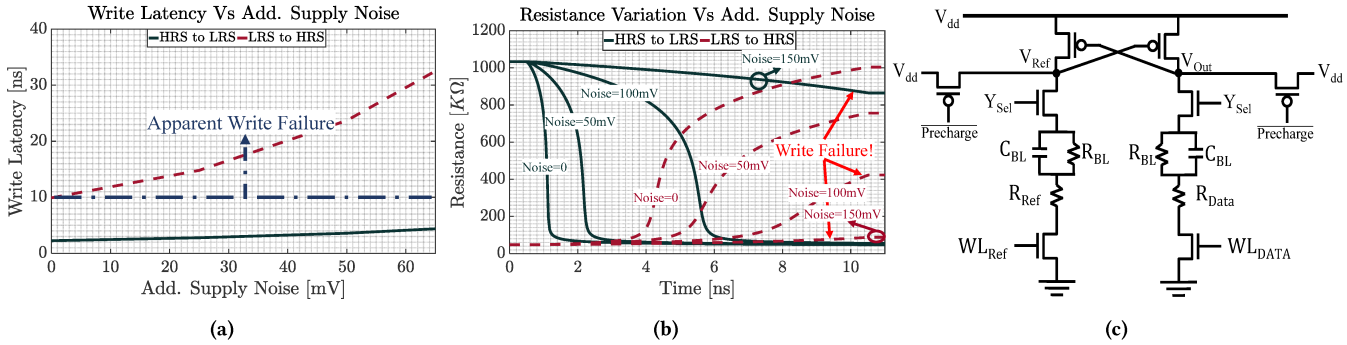


Figure 9: (a) RRAM write latency increases as additional supply noise increases; (b) RRAM resistance variation with additional supply noise; and, (c) single ended read circuitry used it this work (circuit redrawn from US patent no. US7515461B2) [27]. We have considered R_{BL}=25Ω, R_{Ref}=500KΩ and C_{BL}=25fF.

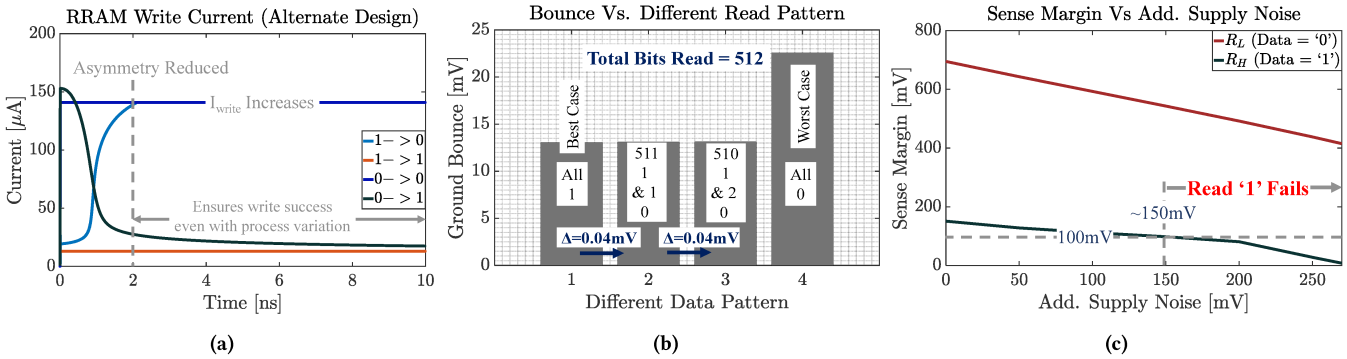


Figure 10: (a) RRAM write current profile for alternate design (symmetric); (b) ground bounce generation vs different read data pattern; (c) sense margin with additional supply noise. Sense margin for data '1' suffers more compared to data '0', and read failure is observed above 150mV of additional supply noise.

read circuitry shown in Fig. 9c). This is true since the final resistance is greater than R_{Ref} (=500KΩ). Therefore, additional supply noise beyond 50mV and less than 120mV will cause LRS to HRS write failure but still can write HRS to LRS successfully. If adversary can generate supply noise in a way that the victim incurs additional supply noise in this range, it will launch $0 \rightarrow 1$ polarity fault injection attack. However, if the victim incurs additional supply noise > 120mV, it will cause complete write failure i.e. DoS attack. Result indicates that adversary can launch DoS and fault injection attack by writing P00 pattern if victim is writing P11 (best-case i.e., lowest self-inflicted noise) at a location with $0 < R_{Int} < 22\Omega$ and $22 < R_{Int} < 37\Omega$ respectively. If the victim generates more self-infected noise, the attacks can cover even larger address space. Note that by preventing $0 \rightarrow 1$ write in victim's memory for several cycles (while allowing $1 \rightarrow 0$ write), the adversary can ensure that most of the plaintext bits will eventually become 0 which can reveal the full key (or partial key).

Controlling polarity of fault injection: The RRAM employed in this work takes longer latency for writing $0 \rightarrow 1$. Therefore, $0 \rightarrow 1$ fault injection is possible with additional supply noise. If the RRAM design takes longer latency for writing $1 \rightarrow 0$, only $1 \rightarrow 0$ fault injection would be possible.

In prior work, designs have been proposed to eliminate the asymmetry using asymmetric doped transistor [28]. Fig. 10a shows the write current profile for all possible four cases of write operation for such alternate (symmetric) design. Write current for $1 \rightarrow 0$ and $0 \rightarrow 0$ increases (more supply noise) although asymmetry is almost eliminated (by reversing TE/BE of RRAM cell, reducing access transistor threshold voltage, V_T by 100mV and using separate write voltages for $V_{1 \rightarrow 0}$=2V and $V_{0 \rightarrow 1}$ = 1.8V). We still kept the write time 10ns to successfully write even with process variation. We investigated this circuit for possible fault injection attack. We have observed for this symmetric design that if the write operation incurs 132mV of noise with a period of 1ns, and the noise continues for entire write operation (10ns), only $1 \rightarrow 0$ write is successful. Furthermore, if the write operation incurs 116mV of bounce with a period of 10ns, only $0 \rightarrow 1$ write is successful. Therefore, symmetric designs could be worse, and both polarity fault injection would be possible. Note that 1ns-periodic-noise can be generated by read operation along with parallel write operation (for higher magnitude). Furthermore, 10ns-periodic-noise noise can be generated by write operation alone.

Detection of victim's write initiation: Adversary needs to know when and where (physical location) the user is writing in order to launch DoS or fault injection attack effectively. One possible

approach adopted by adversary is to store data that generates highest supply noise (i.e., all 0s in the stored data) in various locations of the memory and read them frequently. If a read error occurs, it can be assumed that the victim has initiated a write operation near-by which caused failure in victim's read operation. This is true since victim's read cannot generate enough noise to cause failure to adversary's parallel read operation. The adversary can keep reading many different addresses to detect victim's write operation as read latency is significantly lower than write operation. However, detection of victim's write operation by observing adversary's write failure is not feasible because of long write latency.

3.2 Fault Injection by Read Failure

Fig. 9c presents the single ended read circuity [27] used in this work. The read circuity is proposed in [27]. We have considered R_{BL} = 25Ω, R_{Ref} = 500KΩ and C_{BL} = 25fF for read operation analysis. Fig. 10b shows the supply noise (= ground bounce, droop due to read ignored as mentioned before) generated by various read data pattern. Therefore, the adversary can control the generated supply noise by reading specific data pattern. Adversary can store these data patterns in his memory space before launching the attack.

We further analyze the sense margin for both data '0' and '1' . Fig. 10c shows that sense margin reduces with additional supply noise. However, ground bounce affects sense margin more compared to droop as it, i) reduces the discharge current, and ii) reduces V_{GS} of access transistor (i.e., $R_{Transistor}$ is higher) while voltage droop only reduces the discharge current. It is evident from Fig. 10c that if the adversary can generate supply noise (by write or write + read) in a way that the victim incurs additional supply noise > 150mV, the victim will read data '1' incorrectly. However, sense margin for data '0' is above 150mV even with 350mV of additional supply noise. Therefore, both polarity read failure (DoS by read failure) might not be possible as the required supply noise is too high. We conclude that selection of polarity of fault injection during read operation is not possible with single ended voltage-based sensing as R_L (data '0') will always discharge faster compared to R_H (data '1').

4 DISCUSSION

4.1 Attack Countermeasures

Following techniques can prevent or alleviate the attacks:

- Sequential read/write access: This can be a naïve solution as non-pipelined access hurts system throughput. However, adversary will not be able to create supply noise or inject fault by launching parallel access;
- Intelligent architecture: Parallel operations of different processes can be initiated to addresses with highest possible R_{Int}. This will alleviate the issue to some extent;
- Good quality power/ground grid: A good power/gnd grid reduces R_1 (in Fig. 4) which in turn reduces supply noise. However, this cannot eliminate the issue completely;
- Power rail separation for each bank: Separation of V_{dd} and V_{SS} rails between parallel accessed banks will prevent propagation of supply noise from one bank to another. However, this will incur significant area-overheard and reduce the power rail capacitance (which is not desirable);
- Slow down the system clock: Higher T_{Clock} gives more time to read/write operation at lower headroom voltage to fix latency failures. However, T_{Clock} has to be at least twice (2X throughput loss) to prevent fault injection for just 80mV of additional supply noise (result extended from Fig. 9b).

4.2 Dependency on Memory Technology & Power Grid Design

This study is carried out for RRAM LLC with a specific power grid design. However, we believe that the conclusions drawn in this work are applicable to broad range of NVMs and power grids. A better grid or NVM with scaled read/write current could reduce the amount of supply noise but may not eliminate the challenge completely.

4.3 Memory Testing & Attack Effectiveness

The proposed attacks can be prevented if exhaustive testing of the memory is performed considering all worst-case patterns, all worst-case possible location combinations, all possible parallel read/write conditions (performed under bit-to-bit process variations and wide range of temperatures (-100°C to 200°C for example)) and optimal V_{dd}/T_{Clock} is chosen accordingly. On one hand, exhaustive testing of the memory with fully integrated system is impractical as it: i) increases test time and time to market unacceptably; ii) guard-bands system performance for situations that may never arise under normal workloads (especially true for high performance application).

Let's consider that the noise generated at one address of a particular bank can affect 60% addresses of the nearest bank. Now, one can argue that only the worst-case (nearest of that 60% lowest R_{Int} addresses) can be tested in order to detect error caused by the worst-case supply noise of one address. If write failure is not found, the chip passes the test. Read failure need not be tested as it requires higher noise compared to the required noise for write failure. Firstly, this test has to be done for each of the addresses of all the banks. Such test consumes an extra of 0.68ms [2 parallel-write (10ns) + 1 read (0.5ns, verify) × 1 (only nearest address of other bank) × total number of address)] to test a chip similar to the one employed in this work. Secondly, it is also possible that the other nearest addresses can also get affected (even if the closest one is not) as the bits of those addresses can be significantly weaker due to process variation. Therefore, the test time can be extended (testing all addresses in that range) to 27.06s per chip! This is unacceptable as the total test time (including all other tests) of a chip is typically around 2-3sec (unit test time = 2.7sec shown in [22]). Thirdly, many chips have to be discarded even if a single failure is observed (highly likely) or compromise with the system throughput. Furthermore, adversary can also use ambient temperature in unspecified ranges (i.e. thermal attack) to accelerate the design (for example, F_{clock}) naturally, and force timing failures that were not detected by test. Therefore, adversary can carefully choose patterns, read/write sequences, memory locations and ambient conditions to trigger failures. The weak bits under process and temperature variations will be the likely candidates of such attacks.

5 CONCLUSION

In this work, we show that high write current of NVM can lead to supply noise such as voltage droop and ground bounce. The noise can propagate to the neighboring banks and affect parallel read/write operation. The adversary can control the magnitude of the generated supply noise by various read/write data patterns and leverage this to launch DoS and fault injection attack.

ACKNOWLEDGEMENT

This work is supported by Semiconductor Research Corporation (SRC) (2727.001), National Science Foundation (NSF) (CNS- 1722557, CCF-1718474 and DGE-1723687) and DARPA Young Faculty Award (D15AP00089).

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/wp-content/uploads/RM24C256DS_085.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, 2015. [Online; accessed May-03-2018].

[9] "FM28V102A 1-Mbit (64 K × 16) F-RAM Memory." http://www.cypress.com/file/140901/download, 2015. [Online; accessed May-03-2018].

[10] A. Chen, "A review of emerging non-volatile memory (nvm) technologies and applications," *Solid-State Electronics*, vol. 125, pp. 25 – 38, 2016. Extended papers selected from ESSDERC 2015.

[11] C. J. Xue, G. Sun, Y. Zhang, J. J. Yang, Y. Chen, and H. Li, "Emerging non-volatile memories: Opportunities and challenges," in *2011 Proceedings of the Ninth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, pp. 325–334, Oct 2011.

[12] A. De, M. N. I. Khan, J. Park, and S. Ghosh, "Replacing eflash with sttram in iots: Security challenges and solutions," *Journal of Hardware and Systems Security*, vol. 1, pp. 328–339, Dec 2017.

[13] S. Motaman, M. N. I. Khan, and S. Ghosh, "Novel application of spintronics in computing, sensing, storage and cybersecurity," in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 125–130, March 2018.

[14] J.-W. Jang, J. Park, S. Ghosh, and S. Bhunia, "Self-correcting STTRAM under magnetic field attacks," in *Proceedings of the 52Nd Annual Design Automation Conference*, DAC '15, (New York, NY, USA), pp. 77:1–77:6, ACM, 2015.

[15] S. Ghosh, M. N. I. Khan, A. De, and J. W. Jang, "Security and privacy threats to on-chip Non-Volatile Memories and countermeasures," in *2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–6, Nov 2016.

[16] 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.

[17] M. N. I. Khan, S. Bhasin, A. Yuan, A. Chattopadhyay, and S. Ghosh, "Side-channel attack on STTRAM based cache for cryptographic application," in *2017 IEEE International Conference on Computer Design (ICCD)*, pp. 33–40, Nov 2017.

[18] 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.

[19] M.-J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y.-B. Kim, C.-J. Kim, D. H. Seo, S. Seo, U.-I. Chung, I.-K. Yoo, and K. Kim, "A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures," *Nature Materials*, vol. 10, pp. 625 EP –, Jul 2011. Article.

[20] S. Bhattacharya and D. Mukhopadhyay, "Formal fault analysis of branch predictors: attacking countermeasures of asymmetric key ciphers," *Journal of Cryptographic Engineering*, vol. 7, pp. 299–310, Nov 2017.

[21] Y. Cai, S. Ghose, Y. Luo, K. Mai, O. Mutlu, and E. F. Haratsch, "Vulnerabilities in MLC NAND flash memory programming: Experimental analysis, exploits, and mitigation techniques," in *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 49–60, Feb 2017.

[22] 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.

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

[24] "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].

[25] 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 Conferencee*, pp. 102–104, June 2007.

[26] 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.

[27] T. D. Happ, H. L. Lung, and T. Nirschl, "Current compliant sensing architecture for multilevel phase change memory," 2009. Patent No. US7515461B2, Filed January 5th, 2007, Issued April 7th., 2009.

[28] S. H. Choday, S. K. Gupta, and K. Roy, "Write-optimized STT-MRAM bit-cells using asymmetrically doped transistors," *IEEE Electron Device Letters*, vol. 35, pp. 1100–1102, Nov 2014.