SecNVM: Power Side-Channel Elimination Using On-Chip Capacitors for Highly Secure Emerging NVM

Karthikeyan Nagarajan, *Student Member, IEEE*, Farid Uddin Ahmed, *Student Member, IEEE*, Mohammad Nasim Imtiaz Khan, *Student Member, IEEE*, Asmit De, *Student Member, IEEE*, Masud H. Chowdhury, *Senior Member, IEEE*, and Swaroop Ghosh, *Senior Member, IEEE*

***Abstract*— Emerging nonvolatile memories (NVMs), such as resistive RAM (RRAM) and spin-transfer-torque RAM (STTRAM), present exciting opportunities for data storage applications and offer improved access speeds, retention times, power consumption, and scalability. However, these technologies leak the Hamming weight of data through power side-channel during read and write operations. We propose a technique leveraging on-chip capacitor and voltage regulator (VR) that powers the NVM read/write operations. The side-channel leakage is eliminated due to the isolation of memory array from the external power supply during read/write operations. The residual charge on capacitor bank is discarded safely to prevent information leakage during capacitor recharging. The VR ensures a steady voltage during the entire read/write operations even though the capacitor discharges. The design presents a performance (instructions per cycle) degradation of 0.53%–1.2% under parsec and splash-2 benchmarks and incurs an area overhead of ∼3.54 × 10^{-5}% and an energy overhead of ∼3.05 × 10^{-5}% for a 4-Mb RRAM memory array. For a 64-bit word, the design improves security by 2.7 × 10^{19}× to 2^{64}×. SecNVM should be used in small security-critical memory macros to limit the overhead. SecNVM is generic and could protect any security module such as encryption engines, against power side-channel attacks.**

***Index Terms*— Countermeasures, information leakage, nonvolatile memory (NVM), security, side-channel attack (SCA).**

I. Introduction

TRADITIONAL CMOS-based technologies for data storage applications, such as cache, main memory, and storage, include SRAM and DRAM. However, they store data based on charge, which can leak and, thereby, suffer from a number of issues such as reliability degradation and low noise margin, especially in nanoscaled nodes. Emerging nonvolatile memories (NVMs), such as spin-transfer-torque RAM (STTRAM), resistive RAM (RRAM), and ferroelectric FET (FeFET), present improved speed, power consumption, scalability, and retention times and enable in-memory computing (IMC). While exciting at first glance, they suffer from new sources of variability and security issues. A key problem with emerging NVM is its asymmetric read and write operations. Side-channel emanations, including the power, electromagnetic field, and timing signatures, can reveal the information being written to/read from an NVM bitcell.

RRAM devices suffer from asymmetric read/write currents [1], [2]. STTRAM also suffers from similar issues [3]–[5]. Fig. 1 shows an overview of a power side-channel attack (SCA) leveraging the asymmetric nature of NVM read/write current that can leak sensitive memory data. The adversary extracts the current profile from the V_{DD} pin during the read/write operations on the NVM array. The adversary then analyzes the current profile to extract key features, such as average current, peak current, and access latency during the

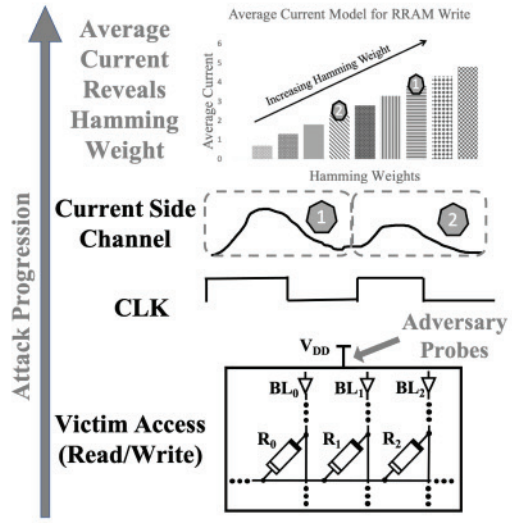


Fig. 1. Threat model: extracting HW of data from RRAM using asymmetric read/write current.

Manuscript received November 7, 2020; revised March 2, 2021 and April 25, 2021; accepted May 16, 2021. Date of publication June 18, 2021; date of current version August 2, 2021. This work was supported in part by Semiconductor Research Corporation (SRC) under Grant 2847.001; and in part by NSF under Grant CNS-1722557, Grant CCF-1718474, Grant DGE-1723687, and Grant DGE-1821766. *(Corresponding author: Karthikeyan Nagarajan.)*

Karthikeyan Nagarajan and Asmit De are with the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802 USA (e-mail: kxn287@psu.edu; asmit@psu.edu).

Farid Uddin Ahmed and Masud H. Chowdhury are with the Department of Electrical Engineering, University of Missouri-Kansas City, Kansas City, MO 64110 USA (e-mail: fahmed@mail.umkc.edu; masud@umkc.edu).

Mohammad Nasim Imtiaz Khan and Swaroop Ghosh are with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802 USA (e-mail: muk392@psu.edu; szg212@psu.edu).

Digital Object Identifier 10.1109/TVLSI.2021.3087734

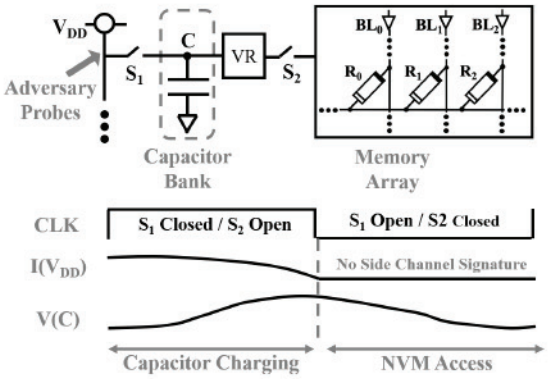


Fig. 2. SecNVM: capacitor-based SCA elimination.

write cycle. Next, the adversary matches the extracted feature (let us say average current) to a previously modeled dataset to match and identify the Hamming weight (HW) of the data accessed. A detailed description of the attack can be found in prior work [6].

A. Related Work

SCA can extract sensitive information through a variety of methods, such as power [7], [8], timing [9], [10], and electromagnetic emanations [11], [12]. Power and timing SCAs are the most predominant. Prior work have proposed methods against side-channel emanations by reducing signal-to-noise ratio (SNR) of the leaked data, power balancing, or by using power isolation. Reducing SNR [13] through noise injection is not an optimum technique. Power balancing techniques include sense-amplifier-based logic (SABL) [14], dual-rail circuits [15], [16], and wave dynamic differential logic (WDDL) [17]. While power balancing eliminates power side channel, it consumes 4× power overhead, 3× area, and 4× performance degradation [17]. Isolation-based techniques include switched capacitor designs [18]–[20] and integrated voltage regulator (VR)-based designs [21]–[24]. The switched capacitor designs incurs 2× performance degradation and 33% power overhead [19]. VR reduces the side channels but fails to completely eliminate it (shown in Section II-B). An adversary with high-resolution measurement tools can still extract sensitive data.

In this work, we propose SecNVM that isolates the power pin from the memory array during read/write operation. The memory is powered through on-chip capacitor banks that act as a battery. This isolation effectively eliminates the side-channel signature observed by the adversary (through V_{DD} pin) and, thereby, prevents power SCA with no performance degradation. Fig. 2 shows the basic idea of SecNVM architecture. A VR is employed to regulate the output of on-chip capacitor during read/write of the memory. During the charging phase, the V_{DD} source is used to charge a capacitor bank (C). The current observed during capacitor charging does not reveal any sensitive information. Once charged, the V_{DD} pin is isolated and the capacitor voltage is fed to a VR, which delivers a constant supply to NVM. Adversary cannot extract side channels since the V_{DD} pin is isolated during the NVM access.

B. Motivation for NVM SCA Elimination

SCA on NVM read/write currents greatly reduces the complexity for the attacker to extract the written/read data. The complexity of brute force attack is 2^n to determine an n-bit data correctly without SCA. However, the adversary can determine the HW (w) of the data (i.e., the number of 1 s in n bits) through SCA. Therefore, the computational effort reduces by a factor of $2^n/\binom{n}{w}$. This is due to the lower number of combinations the adversary has to try since the number of 1s and 0s in the n-bit data is revealed by SCA. Assuming 64-bit data, w can range between 0 and 64. For $w = 0$ or 64, only one combination is possible (all "1"s or all "0"s, respectively). The maximum number of combinations occurs when w is 32. For 64 bits, SCA reduces the computational needs ($2.7 \times 10^{19}\times$ to $2^{64}\times$). This provides a strong motivation to prevent adversary from obtaining fine-grained side channels for sensitive data, such as cryptographic keys and TLB.

C. SecNVM Models

This article presents the following four SecNVM models for NVM side-channel elimination.

Model 1: On-chip VR-based NVM access.
Model 2: On-chip cap (i.e., power bank)-based NVM access.
Model 3: On-chip cap + VR-based NVM access.
Model 4: On-chip cap+VR+charge recycle-based NVM access.

SecNVM can be extended to any NVM that is susceptible to SCA. In summary, the following contributions are made.

1) We propose an on-chip capacitor-based security model to eliminate side-channel leakage in emerging NVM.
2) We propose four power SCA elimination techniques with progressively increasing security for RRAM accesses.
3) We analyze the effectiveness of SecNVM for RRAM and STTRAM.
4) We performed process variation (PV), area, and power analysis.

This article is organized as follows. Section II describes the background on RRAM and STTRAM. Sections III presents the details on SecNVM models and VR. Section IV presents the discussions on the proposed designs. Finally, Section V draws the conclusion.

II. Background on NVM

In this section, we present the basics of RRAM and STTRAM.

A. Basics of RRAM

RRAM contains an oxide material between its top/bottom electrodes (TE/BE) [see Fig. 3(a)]. RRAM resistive switching is due to oxide breakdown and reoxidation, which modifies a conduction filament (CF). The two states of the RRAM are termed low-resistance state (LRS) and high-resistance state (HRS). The process of switching the state to LRS (HRS) is known as SET (RESET). We have used the ASU RRAM

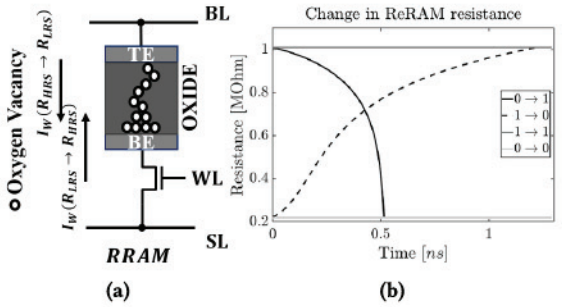


Fig. 3. (a) RRAM bitcell. (b) RRAM resistance switching during all four write operations.

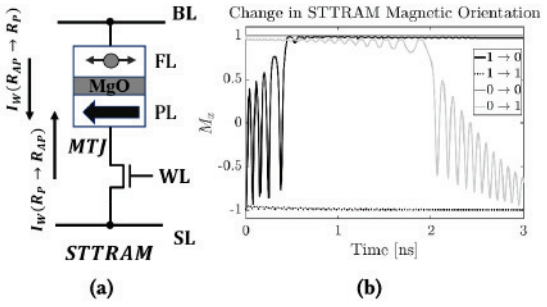


Fig. 4. (a) Schematic of STTRAM bitcell. (b) Switching in the STTRAM magnetic orientation.

TABLE I
RRAM PARAMETERS USED FOR SIMULATIONS

Parameter	Value
Oxide Thickness, RRAM Gap (min/max)	5nm, 1.23nm/1.42nm
RRAM LRS/HRS	0.6 MΩ/1.2 MΩ
Atomic Distance of Oxide	0.25nm
Atomic Energy (Vacancy Gen./Recom)	1.501eV / 1.5eV
V_{write}/V_{read}	2.2V/1V

Verilog-A model [1] along with PTM 65-nm technology for the analysis. The RRAM is bipolar HfO_x-based resistive switching memory [1]. Table I summarizes the model specification extracted from [1]. Note that while this model offers a minimum and maximum RRAM oxide gap of 0.1 and 1.7 nm, we have restricted the min and max gaps of the RRAM employed in this work to 1.23 and 1.42 nm, respectively, in order to reduce the asymmetry in RRAM access currents. Further details regarding this can be found in Section III-A. Change in the min/max gap also affects the read and write voltage of the RRAM bitcell.

Asymmetric Read/Write Current: Fig. 3(b) shows the change in RRAM resistance for all four write combinations. Fig. 5(a) shows the current profile observed in each of these cases. Fig. 5(b) shows the read current profile for data "0" and "1." An adversary can distinguish the four write operations by observing the write current and extract stored data by observing the read current. Furthermore, the adversary can accurately estimate the HW of an n-bit data being written into or read from RRAM by monitoring the current drawn from the power (V_{DD}) pin. Note that adversary can also discriminate the read and write operation, respectively, based on the magnitude and duration of the current drawn by the memory.

B. Basics of STTRAM

Fig. 4(a) shows the STTRAM cell schematic with magnetic tunnel junction (MTJ) as the storage element. MTJ contains a free (FL) and a pinned (PL) magnetic layer. The resistance of MTJ stack is high (low) if FL magnetic orientation is antiparallel (parallel) compared to the PL. The MTJ can be toggled from parallel (P) (data "0") to antiparallel (AP) (data "1") (or vice versa) using current-induced spin torque transfer by passing the appropriate write current from source line to bitline (or vice versa). Fig. 4(b) shows the change in magnetic orientation (M_x) of STTRAM during write operation. Similar to RRAM, STTRAM also shows asymmetric write [Fig. 5(c)] and read [Fig. 5(d)] current.

C. Threat Model

1) Attack Vector Covered in This Article: Fig. 5 shows that RRAM and STTRAM suffer from asymmetric read and write currents. The memory power profile is easily accessible by the adversary since the memory array draws its current from an external power source (i.e., V_{DD} pin). The threat model addressed in this article is adversarial monitoring of the power drawn by using simple probes at the V_{DD} input to steal the plaintext data stored in the NVM. This can be achieved by analyzing the extracted power profiles to infer critical features, such as average current, peak current, or access latency during the read/write cycles (as shown in Fig. 1) that are unique for different HWs of the written/read data. Knowing this HW reduces the computational effort of adversary to extract the plaintext by $\binom{n}{w}$, where n is the size of the word accessed and w is the HW extracted. This threat model has also been explored previously in [6], [25], and [26]. Note that the considered power SCA on NVM is different than conventional power SCA on cryptographic primitives where the objective is to break the key and subsequently steal the plaintext data sent over the network. In order to better understand the scope of this article, the memory hierarchy and the critical secured and unsecured components of a system are listed in the following.

1) *Memory Heirarchy:* The system considered for our attack consists of: 1) CPU with last level cache (LLC) as NVM-based; 2) main memory (DRAM); and 3) hard disk (NVM).
2) *Unsecured Components:* The following components are assumed unsecured: 1) unencrypted NVM data (cache) and 2) power pins of the chip/board/device during memory read/write operation. This is possible if adversary has access to the device or the power port.
3) *Secure Components:* The following components are assumed secured: 1) encrypted NVM data (hard disk); 2) internal nodes within the chip, e.g., NVM array; and 3) system bus, e.g., command/data bus.
4) *Attack Scenarios:* 1) Insider adversary (e.g., adversary in cloud computing farm who can place a probe in the power port and collect the traces) and 2) physical possession/attack by the user: malicious insertion of

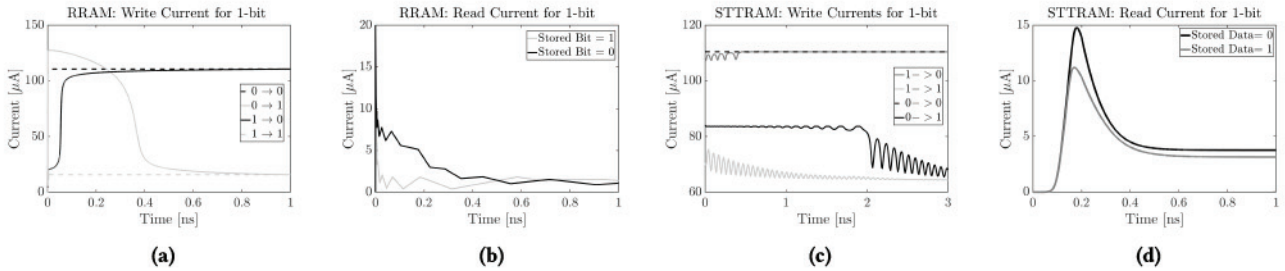


Fig. 5. (a) RRAM write current profiles for four write combinations. (b) RRAM read current profile for data 0 and 1. (c) STTRAM write current profiles for all four write combinations. (d) STTRAM read current profile for data 0 and 1.

power logger and transmitter to devices such as PCs, gaming consoles, and power adapters during use by adversary who has physical access to device (e.g., public computer). Finally, academic researchers/white hat adversaries who aim to investigate security vulnerabilities of systems to develop countermeasures can also conduct this experiment.

2) Attack Vectors Not Covered in This Article: This includes: 1) power-monitoring and timing SCA on cryptographic operations; 2) electromagnetic SCA; 3) timing SCA on cache; and 4) monitoring the external data bus to CPU.

D. SCA Vulnerability Evaluation Criteria

NVM asymmetric read and write current profiles need to be masked to protect it from SCA. The proposed SecNVM achieves this objective. The adversary observes a constant current drawn by the memory irrespective of write data polarity and old data stored in the cell for write operation and read data. The following considerations are made in this work.

1) Write Operation: For each of the four write combinations, we consider 8 bits are written together. This maximizes the difference in the write current profile (if there is any) and facilitates better detection. Next, we tap the V_{DD} pin for each of the SecNVM models to monitor the current and to identify the write combination. For a completely secure design, all four combinations should provide the exact same current profile at V_{DD} pin. The adversary can still launch the attack successfully if a model minimizes the difference but not eliminate it completely. It will require very sophisticated and expensive measuring tools (to detect any small variation).

2) Read Operation: The considerations are similar as write except that there are two current profiles during read (for data 0 and 1).

III. SecNVM Design

In this section, we present the SecNVM models that progressively improve NVM security by taking RRAM as a test case.

A. Simulation Setup

The RRAM model described in Section II-A is used as a representative NVM to depict the effectiveness of the proposed SecNVM models. While memory architectures may widely vary on the number of bits being written/read in parallel,

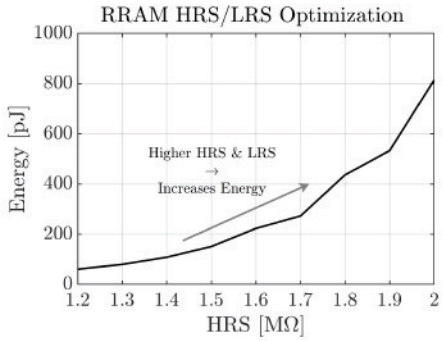


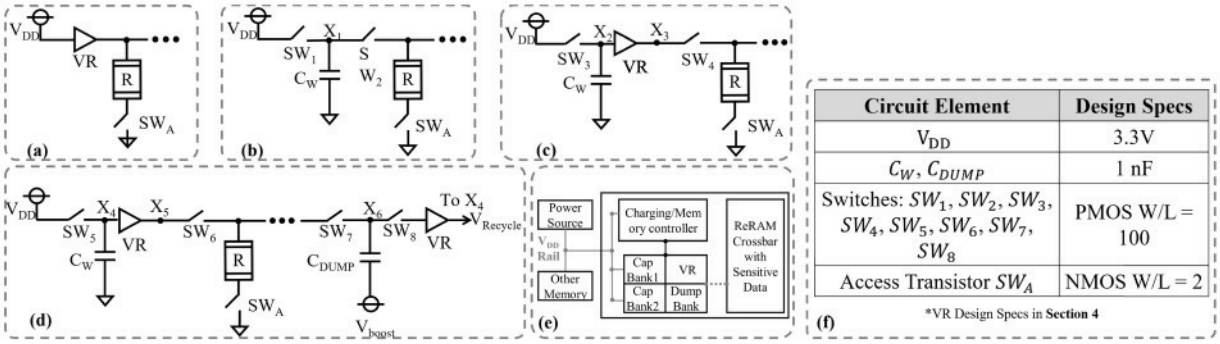
Fig. 6. Impact of HRS/LRS on RRAM write energy.

we have chosen 8 bits as a representative example. We have modeled the current drawn by eight RRAM cells during write and read operations. We also consider a V_{DD} source of 3.3 V and the read/write voltage provided to the RRAM array is 2.2 V. The write and read operations on the RRAM cells are then conducted under each of the SecNVM models and the corresponding side-channel currents (i.e., current observed at the V_{DD} pin) are extracted. Note that in order to avoid read disturb, the read current of the RRAM array is reduced by applying low read voltage (a few hundred mV) using a clamp circuit. This access circuitry is a part of the RRAM array design. For our experimental purposes, we have considered the RRAM memory array as an isolated system and not as a part of an SoC.

RRAM HRS/LRS Optimization: The selection of HRS and LRS values for the RRAM determines the write latency, the read/ write current drawn, and subsequently the side-channel signature at the V_{DD} pin. Ideally, we aim to reduce the average energy consumed by the RRAM accesses and reduce the asymmetry in read/write currents. We set the tunneling magnetoresistance (TMR) (= (HRS − LRS)/LRS) to be 1. Fig. 6 shows that lower HRS and LRS values consume much lower write energy (= average power × write latency) per RRAM access. This is primarily due to faster latency compared to high HRS/LRS values. To reduce the load on the proposed capacitors, we have considered HRS/LRS = 1.2 MΩ/0.6 MΩ.

B. Model 1

In model 1 [Fig. 7(a)], the array is operated by the power supply through a VR (details in Section III-G), which uses



Circuit Element	Design Specs
V_{DD}	3.3V
C_W, C_{DUMP}	1 nF
Switches: SW_1, SW_2, SW_3, SW_4, SW_5, SW_6, SW_7, SW_8	PMOS W/L = 100
Access Transistor SW_A	NMOS W/L = 2

Fig. 7. SecNVM models. (a) Model 1: only VR. (b) Model 2: capacitor bank. (c) Model 3: capacitor Bank + VR. (d) Model 4: capacitor Bank + VR + charge recycling. (e) Chip-level schematic of SecNVM showing V_{DD} rail (in gray) and internal rail to power the NVM (in gray dotted line). (f) Design specifications for the SecNVM models.

V_{DD} of 3.3 V and delivers a constant 2.2 V to the RRAM. Fig. 8(a) shows the current drawn by RRAM for 8-bit write operation (top subplot) and the current observed by adversary at the V_{DD} pin (bottom subplot). Results indicate that all four current profiles can be distinguished with a large margin (min resolution = 10 μA). A similar conclusion can be drawn for read operation. Note that model 1 is the traditional design and the base case for this work.

C. Model 2

Since model 1 does not offer adequate protection against SCA, we propose model 2 that incorporates a 1-nF capacitor to power the RRAM [see Fig. 7(b)].

Initially (i.e., at boot-up), switch SW_1 is closed and SW_2 is open, and V_{DD} charges the capacitor (C_W) to 2.2 V. During this time, no write/read occurs in the memory array since they are isolated from the power source due to the open SW_2. Next, SW_1 is opened and SW_2 is closed. The charge stored at C_W is used to write/read the array. The current consumption during write operations results in discharging of node X_1 at a rate that depends on the HW of the write data. The minimum X_1 node voltage is set to 1.7 V (lower threshold for RRAM access). Simulation results show that the RRAM cells are not accurately/completely written/read under 1.7 V.

In the following cycle, the capacitor is recharged (SW_2 is open and SW_1 is closed) and the process is repeated. While model 2 prevents the adversary from observing any real-time RRAM write current profiles, it presents some key security issues.

1) Issue 1: If the capacitor is recharged after a fixed number of cycles (n), the adversary can leverage the current signature during the recharge phase and determine the residual charge in the capacitor after n cycles. Fig. 8(b) shows the write current from the capacitor for writing all four combinations (top subplot) and the corresponding recharge current drawn from the V_{DD} (bottom subplot). Results indicate that $0 \rightarrow 1$ and $1 \rightarrow 1$ pair and $1 \rightarrow 0$ and $0 \rightarrow 0$ pair have almost a similar profile with variation less than 1 μA. The difference between these two pairs is 2 μA.

2) Issue 2: If the capacitor is recharged only after the node reaches 1.7 V (achieved using a comparator), the frequency of the recharging phase observed can reveal coarse-grained information to the adversary. The adversary can discern that 0.5 V of node voltage is discharged after a number of write cycles (let us say m cycles). This again reveals the HW of the data written during the m cycles.

3) Issue 3: Write voltage degrades over the course of the write cycles ($2.2 \rightarrow 1.7$ V), which increases the write time and leads to nonuniform write latency. A variation in write latency beyond allocated time leads to malfunction since memory datapath is pipelined.

D. Model 3

In order to eliminate the security issues from model 2, the recharging mechanism is modified and a VR (details in Section III-G) is added in model 3 [see Fig. 7(c)]. V_{DD} = 3.3 V is used to charge the capacitor node (X_2) to 3.2 V. During the write phase, X_2 is fed into the VR, and the VR outputs a steady write voltage of 2.2 V to the RRAM.

To resolve issue 2 of model 2, the number of write cycles is fixed (let us say i) before recharging and any unused charge (above 2.2 V) in the capacitor is dumped to the GND before the next recharge cycle. This means that after every i cycles of write operation, the capacitor is always discharged to 2.2 V. Therefore, the recharge current profile is the same regardless of the written data in the previous i cycles. Since the frequency of the recharge operation and the current observed is constant, the adversary is unable to extract any information from the write current.

Fig. 8(c) shows the current profiles drawn by RRAM for writing all four combinations of data (top subplot) and the current observed at the V_{DD} pin during recharging the capacitor (bottom subplot). We observe that the recharge current of all four cases is the same. A similar conclusion can be drawn for read operation [Fig. 8(d)] and for any number of read/write operation which totals to i cycles. However, model 3 presents two key security issues.

1) Issue 1: Adversary can still monitor the GND pin and extract the current discharged by the capacitor. This essentially provides coarse-grained information of the written data. Assuming that i writing cycles discharge the write capacitor voltage from 3.2 to 2.8 V, model 3 dumps 2.8–2.2 V = 0.6 V equivalent charge to the GND. The adversary can determine

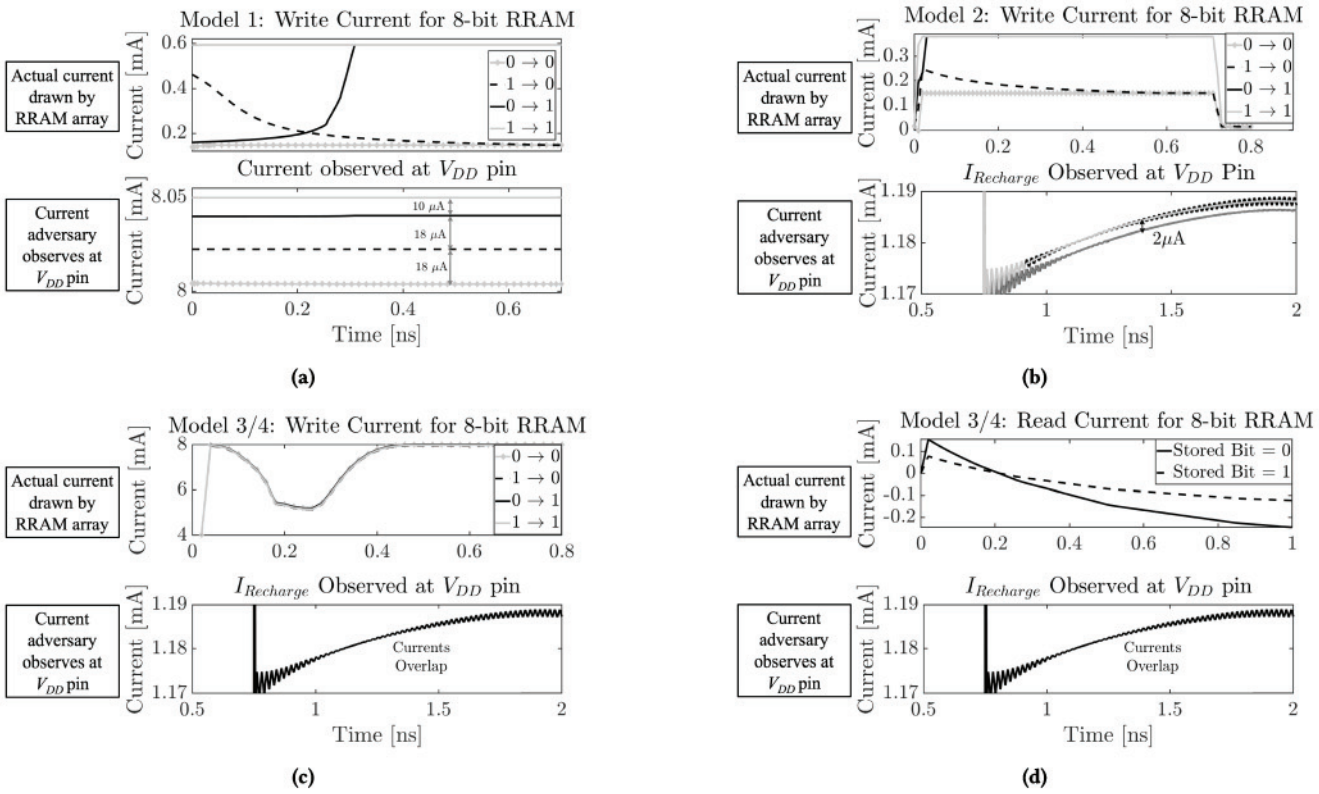


Fig. 8. Current observed at the V_{DD} pin during RRAM write operation with (a) model 1, (b) model 2, (c) model 3/4, and (d) RRAM read operation for model 3/4.

the charge consumed in i cycles and extract the HW of the bits written.

2) Issue 2: Dumping the remaining charge to GND repeatedly after a number of cycles leads to huge energy penalty.

E. Model 4

In order to address the model 3 issues, we introduce an additional capacitor (named C_{DUMP}) in model 4. Furthermore, the VR is also completely powered by the write capacitor (C_W) to completely isolate the write process from the V_{DD} pin [see Fig. 7(d)].

After the fixed number of writes, the residual charge (above 2.2 V) present in C_W is discharged to C_{DUMP}. This ensures that the adversary is unable to monitor the current at the GND pin to determine the unused charge. Note that C_{DUMP} can charge to a maximum of 2.2 V. Initially, V_{boost} is set to GND. Once the charge at node X_6 is saturated (voltage reaches 2.2 V), C_{DUMP} becomes unusable until the charge is recycled back into C_W. We change the V_{boost} value to 1 V to boost the X_6 node voltage to 3.2 V. This node is then used to charge the X_4 node back to 3.2 V. The simulation result of write and read operation for model 4 is the same as in Fig. 8(c) and (d), respectively. Therefore, they are omitted.

In summary, model 4 isolates the V_{DD} pin from the RRAM array and prevents the adversary from directly observing the read/write currents. It incorporates write capacitors (C_W) to store charge and power the RRAM accesses. It also introduces a VR to ensure that a steady voltage is supplied to the RRAM array. In order to prevent the attacker from gaining any information from the frequency of C_W charging, the design recharges C_W after a fixed number of read/write cycles. Excessive charge (>lower threshold) before recharging is moved to the dump capacitors (C_D). Once completely charged, the charge from the dump capacitors is recycled back into the write capacitors to improve the power efficiency of the proposed model. Not dumping excessive charge into the adversary-accessible GND pin also improves security.

F. Pipelined Operation of Models 3 and 4

Models 3 and 4, shown in Fig. 7(c) and (d), respectively, require careful pipelining of the multiple stages to allow uninterrupted memory operation during: 1) charging of the write capacitors; 2) writing of data in the array; 3) dumping of the residual charge into dump capacitors (for model 4 only); and 4) recharging of the write capacitors. The memory controller ensures a pipelined execution of each of these four steps.

We also propose using multiple write and dump capacitors to eliminate the performance degradation during recharging of the capacitor. The design requires three write capacitors, each assigned to either *WRITE*, *STANDBY*, or *CHARGE* by the memory controller. Similarly, the three dump capacitors are needed, which are assigned to either *RECEIVE*, *STANDBY*, or *RECYCLE*. The three write capacitors and three dump capacitors together ensure that a capacitor is always available to power the memory (i.e., no performance penalty).

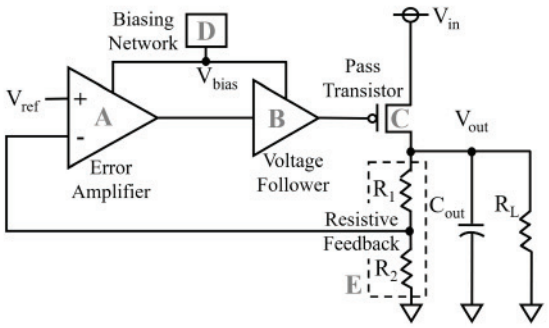


Fig. 9. Block diagram of the LDO used in this work.

Bits Written per Pipeline Cycle: Writing 8 bits in parallel requires a maximum average current of 0.4 mA (for $0 \rightarrow 1$). For a constant load drawing 0.4 mA, simulation shows that voltage at X_4 [in Fig. 7(d)] discharges from 3.2 V to the lower threshold value of 2.2 V after 130 ns. Similarly, we observe that the time required to recharge the write capacitor node (X_4) from 2.2 to 3.2 V is 130 ns.

The write latency for RRAM is 0.7 ns (including a 20% safety margin). Assuming that 8 bits are written at a time, a total of 1480 bits can be written by a 1-nF write capacitor before it requires recharging.

G. VR Design

The traditional low dropout regulator (LDO) consists of an error amplifier, a pass transistor, and a resistive feedback network. The LDO used in this work also contains a voltage follower and a biasing network (Fig. 9) to make it more immune to noise and to ensure a stable and accurate output.

1) Error Amplifier: It is a cascaded structure of a differential amplifier and source amplifier (Block A in Fig. 9) and the main component of the feedback system. The overall gain and phase margin of error amplifier are 66.65 dB and 68.12°, respectively. A 0.5-pF capacitor is used between the first and second stages as a miller capacitor to separate the dominant and nondominant pole for stability purpose.

2) Voltage Follower: The voltage follower (Block B in Fig. 9) acts as a buffer amplifier to create an isolation between the input of the pass transistor and the output of the error amplifier.

3) Pass Transistor: The PMOS transistor (Block C in Fig. 9) is the main driver switch, which maintains the output voltage and delivers current to the load.

4) Biasing Network: It (Block D in Fig. 9) delivers a fixed voltage to the error amplifier and the voltage follower to reduce the effects of input voltage/temperature variation and noise in the supply.

5) Resistive Network: It consists of two resistors R_1 and R_2 (Block E in Fig. 9) and divides the output voltage in half, which is fed to the negative input of error amplifier for correction in voltage output.

C_{out} (Fig. 9) denotes the output capacitor, which stabilizes the output for sudden load variation. R_L presents the load, which is memory array for our work. The design is simulated in Cadence Virtuoso using the 180-nm CMOS technology.

TABLE II
SPECIFICATIONS OF THE LDO USED IN THIS WORK

Parameter	Specification
Voltage (Input/Output), Overhead	(2.4V-4V/2.2V), 50mV
Transient Response, Line Regulation	<10μs, 1.85 (mv/V)
Load Regulation	0.0017 (mV/mA)
Load Current, C_{out}	1mA (max), 1nF

The overall specification is summarized in Table II. The LDO output is 2.2 V for wide range of load with the operating input range of 2.4–4 V.

IV. DISCUSSIONS AND LIMITATIONS

A. Consideration to Other NVMs

We have shown the effectiveness of SecNVM in eliminating side-channel signatures of NVM taking RRAM read and write operations as examples. However, SecNVM can also be extended to other emerging NVMs as well that incurs asymmetric read/write current. Fig. 5(c) shows the current profiles drawn by the STTRAM for all four combinations of write (top subplot) and the current observed at the V_{DD} pin during recharging of the capacitor (bottom subplot). We observe that the recharge current is the same for all four cases. Similarly, Fig. 10(b) shows that the STTRAM read operation also leaks no side-channel signature under model 3/4.

B. Impact of PV

PVs in the RRAM can cause an increase or decrease in this write latency and can alter the capacitor design requirements. We conducted the PV analysis to detect worst case corners where the write latency can be high enough to prevent accurate write operation within the allowed write time. A 100-point Monte Carlo analysis with 3σ of 5% of the RRAM initial resistance with a mean of 1.2 MΩ as HRS and 0.6 MΩ as LRS at $T = 25$ °C shows that the mean write latency is 548 ps and the standard deviation is 24 ps (see Fig. 11). Note that the choice of our mean HRS/LRS values is explained in Section III-A. Assuming a four-sigma variation for 2-kB memory, an additional 96 ps is required (i.e., total time = 644 ps). Therefore, we provide 700 ps of write time that is more than the four-sigma minimum requirement. Note that the maximum latency case of 584 ps is less than our designated 700 ps.

C. Placement of SecNVM Capacitor Bank

The capacitors employed in this work are realized using metal–insulator–metal (MIM) structures embedded in the higher metal layers. The upper metal layers in NVM remain relatively unoccupied (except routing of global/high-frequency signals) since NVMs employ via spacing for the storage element. The MIM capacitor should be designed above the array (or other areas of chip with unoccupied higher metal layers) to hide the area overhead. The designer can ensure that these areas belong to filler cells that do not implement functional logic. Furthermore, design capacitors can be implemented over

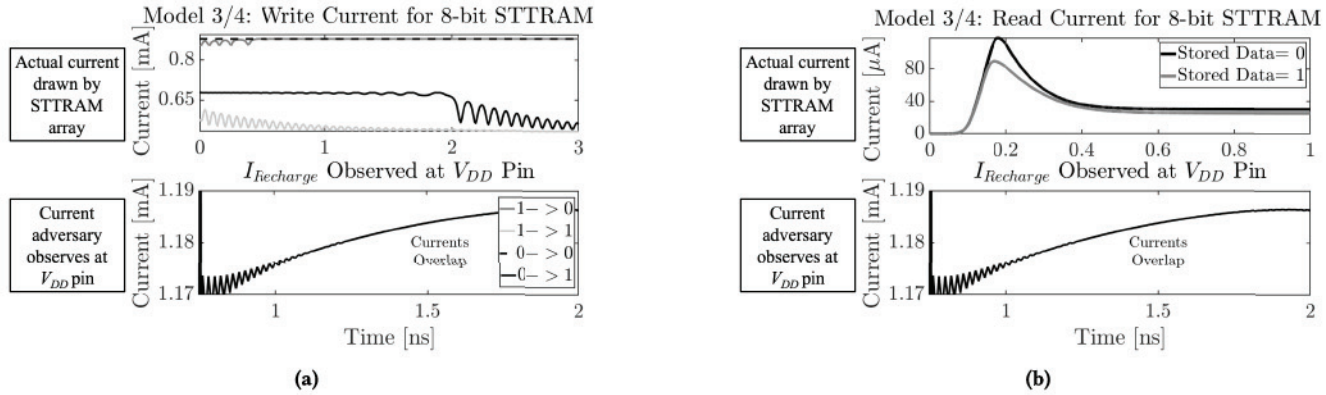


Fig. 10. Current observed at V_{DD} pin under model 3/4 during STTRAM. (a) Write operation. (b) Read operation.

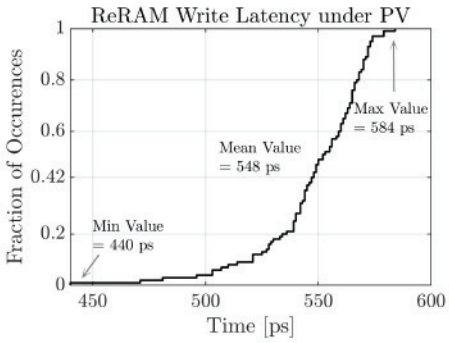


Fig. 11. Impact of PV on the RRAM write latency.

areas that do not contain power-hungry components in the base layers. Note that, the power and clock grid would still be present in the lower level metals. Therefore, the proposed approach will only weaken the power/clock grid locally. The proposed approach does not degrade the metal densities and/or violate any DRC rule. Note that the usage of MIM capacitors realized at higher layer metals is common in analog designs where linear capacitors are highly desirable.

D. Area, Power, and Performance Analysis

The area of a 1-nF capacitor and the VR are 0.5 and 0.7 mm^2 (estimated using TSMC PDK for 180-nm technology node), respectively. Each switch (1–8) consumes 0.42 μm^2. Each of these values was determined from their layout. The area of the VR, designed in a 180-nm technology node, has been scaled down to the 65-nm technology node for area analysis. The estimated area of the VR, after scaling, in 65-nm technology node is 0.18 mm^2. As mentioned in Section IV-C, the capacitors are realized using upper metal layers, so they do not present any silicon area overhead. In addition, VRs may be a part of traditional chips already, and in those cases, they may not incur any area overhead. In our area analysis (Table III), we have presented calculations for both area with and without VR overhead. Model 1 (our baseline model) does not incur any area overhead without VR overhead. Note that the components in each of the models differ. For example, model 1 has only one VR, model 2 has one write capacitor, model 3 has three write capacitors and one VR, and finally, model 4 has three write capacitors, three dump capacitors, and one VR.

TABLE III
COMPARISON OF SECNVM MODELS

Name	Area (mm^2) (with VR)	Area (um^2) (No VR)	Power mW	Elim. SCA?	Energy Efficient?
Model 1	0.18	-	27	No	No
Model 2	8.4×10^{-7}	0.84	0.02	No	No
Model 3	0.18	0.84	59.4	No	No
Model 4	0.36	1.68	61.3	Yes	Yes

The maximum total area of SecNVM model is 0.36 mm^2 and the maximum total power consumed is 61.3 mW (max for model 4 with added VR). Note that the proposed SecNVM design can be used for the entire memory array. As mentioned in Section III-F, the design can read/write 1480 bits at a time with no performance penalty. It takes 130 ns for the proposed 1-nF capacitor to be completely charged. During the recharge phase, the array operations are powered using the STANDBY capacitor. In order to minimize performance penalties for large memory arrays, the proposed WRITE capacitor bank can be divided into smaller capacitors that can each serve multiple arrays. While this increases the recharging frequency, it minimizes read/write latency overhead. For example, a 1-nF write capacitor is divided into four 0.25-nF capacitors to serve different arrays. Each 0.25-nF capacitor can read/write up to 370 bits before recharging at the cost of 37.5 ns (=130 ns/4) latency overhead between writes/reads. Therefore, the proposed on-chip SCA elimination technique can be shared in the whole memory array at the cost of performance penalty. This penalty can be reduced by employing multiple copies of the proposed circuitry at the cost of more area overhead as illustrated in the following. Note that in scaled technologies, the read/write energy of the NVMs is expected to scale down due to material- and device-level innovations. Therefore, the load on VR is expected to reduce in smaller technology nodes. This will translate to small-size capacitors and/or capability to read/write wider memory word sizes with the same capacitor size.

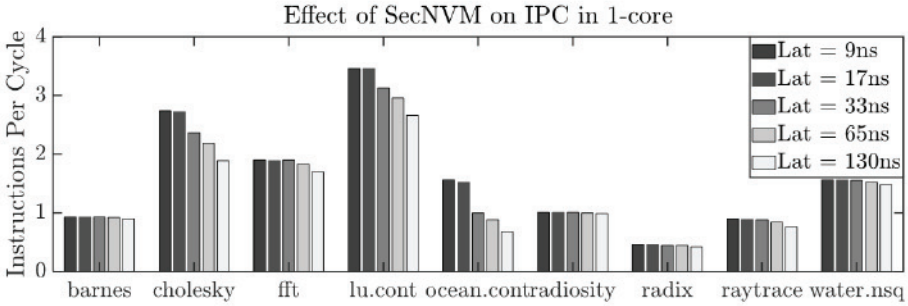


Fig. 12. Single-core system performance under SecNVM for splash-2 benchmarks.

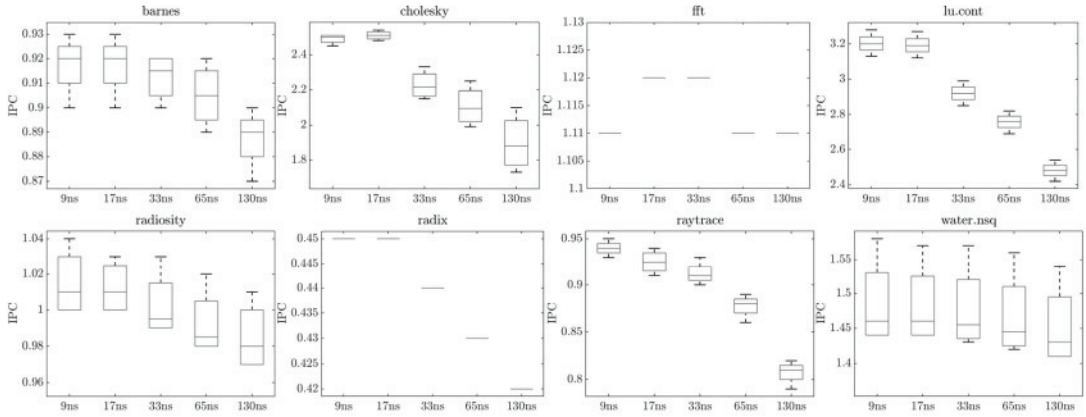


Fig. 13. Quad-core system performance system performance under SecNVM for splash-2 benchmarks.

1) Exemplary Area and Power Overhead: Multiple prior works have determined area and array density of NVM arrays. In [27], a 4-Mb RRAM macro with a test chip area of 4.74 mm^2, implemented in 65-nm CMOS technology, is presented. The design includes a charge pump that regulates the voltage and, therefore, requires no additional VR. Model 4 (without VR) would incur an area overhead of $3.54 \times 10^{-5}\%$ when incorporated within this RRAM macro. The testchip read energy over a read-access time of 45 ns is presented to be 11.2 pJ/bit. Each subarray consists of 1024 columns and 512 rows. Assuming that the SecNVM Model-4 powers one subarray at a time, the total energy overhead incurred is $3.05 \times 10^{-5}\%$. Here, since the RRAM testchip energy includes the VR, we have calculated the energy overhead caused by Model-4 with no included VR. Note that the overhead percentages mentioned are for the baseline case where we incur no performance penalty. This overhead can be decreased if SecNVM resources are shared in a large memory chip among multiple arrays. For example, for a performance penalty of 37.5 ns, the SecNVM resources (i.e., read/write capacitance) will be divided by 4. This will reduce the SecNVM area overhead to $8.85 \times 10^{-6}\%$ and its energy overhead to $7.62 \times 10^{-6}\%$.

2) System-Level Simulations: In order to determine the effect of SecNVM implementation on system-level performance, we ran splash2 benchmarks using SNIPER [28] with a one-core and a quad-core Gainestown (Xeon) CPU clocked at 2.66 GHz. Different performance penalties correspond to different ways in which the SecNVM capacitor may be divided between different memory resources. An access latency of 130 ns corresponds to the time taken to completely charge a 1-nF capacitor that is not divided. If the 1-nF capacitor is divided into 16 smaller capacitors (each 62.5 pF) to serve different areas of the memory array in parallel, an access latency of 9 ns is observed.

We consider all possible SecNVM capacitor divisions that cause performance penalties in the range of 9–130 ns in our system-level implementation. In Fig. 12, the instructions per cycle (IPC) observed for different workloads for each of the possible latencies for a one-core system is shown. Similarly, Figs. 13 and 14 show the IPC range observed in a quad-core system under different SecNVM configurations for splash-2 and parsec benchmarks, respectively. It is seen that the IPC is higher for lower access latencies for each workload. Therefore, it is beneficial to divide the write capacitor bank into smaller capacitors in order to improve IPC. In the quad-core system, under different SecNVM configurations, the IPC drop observed when the R/W latency increases by 1344% (increases from 9 to 130 ns) ranges from 0.6% to 18.6% for parsec benchmarks and from 0% to 30.5% for splash-2 benchmarks. The design presents a tradeoff surface between recharging frequency and IPC degradation. In our example, an optimal SecNVM configuration would be four 0.25-nF capacitors with a recharging latency of 33 ns and an average IPC degradation of 0.53% and 1.2% for parsec and splash-2 benchmarks, respectively.

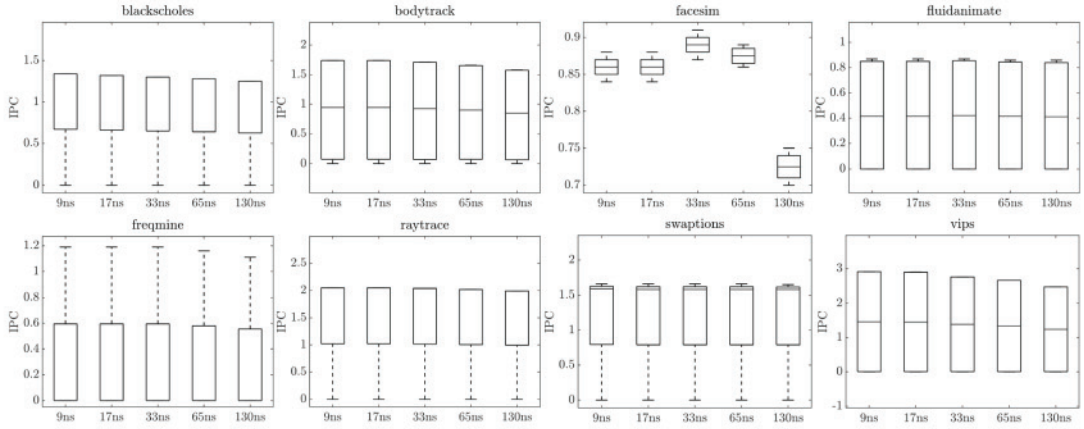


Fig. 14. Quad-core system performance system performance under SecNVM for parsec benchmarks.

E. Application of SecNVM

The proposed SecNVM provides the highest resilience to power SCA at the cost of design overheads. It should be used in security-sensitive memory macros to keep the overhead low. For example, some memory arrays could be dedicated for storing security-sensitive data, e.g., cryptographic keys (both intermediate and final) protected using secNVM. Note that the proposed methodology is generic in nature and can also be extended to isolate the side-channel leakage of cryptographic primitive, such as encryption engines and hash functions. Note that computing systems may employ encryption to ensure data confidentiality. For example, Intel has released total memory encryption [29] that encrypts all memory accessed from the CPU. However, this does not include cache memory that is a part of the CPU. Therefore, the data inside the SoC (in caches) remain plain text. Similarly, AMD offers secure memory encryption (SME) [30]. However, it is used to protect the contents of the MAIN memory (DRAM) from physical attacks on the system. Therefore, it does not address attacks on the cache. The proposed power side-channel extraction attack by an adversary focuses on cache data that are commonly stored in the NVM as plaintext. Attacks on NVM-based caches have been demonstrated in prior work [6], [31], [32], but no research exists on power SCA attacks in NVM-cache applications.

V. Conclusion

Emerging NVMs leak HW of data through power side-channel during read and write operations. We proposed SecNVM, an on-chip capacitor-based technique to isolate the NVM from the power pin during read and write operation. It deters the adversary to exploit the asymmetric NVM current at the power pin to launch SCA.

References

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

[2] S. Yu, X. Guan, and H.-S.-P. Wong, "On the switching parameter variation of metal oxide RRAM—Part II: Model corroboration and device design strategy," *IEEE Trans. Electron Devices*, vol. 59, no. 4, pp. 1183–1188, Apr. 2012.

[3] A. Iyengar, S. Ghosh, N. Rathi, and H. Naeimi, "Side channel attacks on STTRAM and low-overhead countermeasures," in *Proc. IEEE Int. Symp. Defect Fault Tolerance VLSI Nanotechnol. Syst. (DFT)*, Sep. 2016, pp. 141–146.

[4] X. Fong, S. Choday, and K. Roy, "Design and optimization of spin-transfer torque MRAMs," in *More Than Moore Technologies for Next Generation Computer Design*. Springer, 2015, pp. 49–72.

[5] D. Lee, S. K. Gupta, and K. Roy, "High-performance low-energy STT MRAM based on balanced write scheme," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2012, pp. 9–14.

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

[7] P. C. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *Proc. 19th Annu. Int. Cryptol. Conf. Adv. Cryptol.* Berlin, Germany: Springer-Verlag, 1999, pp. 388–397.

[8] E. Brier, C. Clavier, and O. Francis, "Optimal statistical power analysis," Ph.D. dissertation, IACR Cryptol. ePrint Archive, Dept. Secur. Technol., Gemplus Card Int., France, 2003.

[9] P. C. Kocher, "Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems," in *Advances in Cryptology*, N. Koblitz, ed. Berlin, Germany: Springer, 1996, pp. 104–113.

[10] D. Brumley and D. Boneh, "Remote timing attacks are practical," *Comput. Netw.*, vol. 48, no. 5, pp. 701–716, Aug. 2005.

[11] J.-J. Quisquater and D. Samyde, "Electromagnetic analysis (EMA): Measures and counter-measures for smart cards," in *Proc. Int. Conf. Res. Smart Cards*. Springer, 2001, pp. 200–210.

[12] K. Gandolfi, C. Mourtel, and F. Olivier, "Electromagnetic analysis: Concrete results," in *Proc. Int. Workshop Cryptograph. Hardw. Embedded Syst.* Paris, France: Springer, 2001, pp. 251–261.

[13] T. Güneysu and A. Moradi, "Generic side-channel countermeasures for reconfigurable devices," in *Proc. Int. Workshop Cryptograph. Hardw. Embedded Syst.* Nara, Japan: Springer, 2011, pp. 33–48.

[14] K. Tiri, M. Akmal, and I. Verbauwhede, "A dynamic and differential CMOS logic with signal independent power consumption to withstand differential power analysis on smart cards," in *Proc. 28th Eur. Solid-State Circuits Conf.*, Sep. 2002, pp. 403–406.

[15] M. Bucci, L. Giancane, R. Luzzi, and A. Trifiletti, "Three-phase dual-rail pre-charge logic," in *Proc. Int. Workshop Cryptograph. Hardw. Embedded Syst.* Yokohama, Japan: Springer, 2006, pp. 232–241.

[16] D. Sokolov, J. Murphy, A. Bystrov, and A. Yakovlev, "Design and analysis of dual-rail circuits for security applications," *IEEE Trans. Comput.*, vol. 54, no. 4, pp. 449–460, Apr. 2005.

[17] D. D. Hwang *et al.*, "AES-based security coprocessor IC in 0.18-μm CMOS with resistance to differential power analysis side-channel attacks," *IEEE J. Solid-State Circuits*, vol. 41, no. 4, pp. 781–792, Apr. 2006.

[18] A. Shamir, "Protecting smart cards from passive power analysis with detached power supplies," in *Cryptographic Hardware and Embedded Systems*, Ç. K. Koç and C. Paar, eds. Berlin, Germany: Springer, 2000, pp. 71–77.

[19] C. Tokunaga and D. Blaauw, "Securing encryption systems with a switched capacitor current equalizer," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 23–31, Jan. 2010.

[20] P. Corsonello, S. Perri, and M. Margala, "An integrated countermeasure against differential power analysis for secure smart-cards," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2006, p. 4.

[21] M. Kar, D. Lie, M. Wolf, V. De, and S. Mukhopadhyay, "Impact of inductive integrated voltage regulator on the power attack vulnerability of encryption engines: A simulation study," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2014, pp. 1–4.

[22] M. Kar, A. Singh, S. Mathew, A. Rajan, V. De, and S. Mukhopadhyay, "Exploiting fully integrated inductive voltage regulators to improve side channel resistance of encryption engines," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 2016, pp. 130–135.

[23] A. Singh, M. Kar, J. H. Ko, and S. Mukhopadhyay, "Exploring power attack protection of resource constrained encryption engines using integrated low-drop-out regulators," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2015, pp. 134–139.

[24] A. Singh, M. Kar, A. Rajan, V. De, and S. Mukhopadhyay, "Integrated all-digital low-dropout regulator as a countermeasure to power attack in encryption engines," in *Proc. IEEE Int. Symp. Hardw. Oriented Secur. Trust (HOST)*, May 2016, pp. 145–148.

[25] S. Swami and K. Mohanram, "ARSENAL: Architecture for secure non-volatile memories," *IEEE Comput. Archit. Lett.*, vol. 17, no. 2, pp. 192–196, Jul. 2018.

[26] S. Swami, J. Rakshit, and K. Mohanram, "SECRET: Smartly EnCRypted energy efficient non-volatile memories," in *Proc. 53rd Annu. Design Automat. Conf.*, Jun. 2016, pp. 1–6.

[27] M.-F. Chang *et al.*, "A 0.5 V 4 Mb logic-process compatible embedded resistive RAM (ReRAM) in 65 nm CMOS using low-voltage current-mode sensing scheme with 45 ns random read time," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2012, pp. 434–436.

[28] T. E. Carlson, W. Heirman, and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal. (SC)*, New York, NY, USA, Nov. 2011, pp. 1–12.

[29] Intel. (2021). *Intel Architecture Memory Encryption Technologies*. Accessed: Apr. 2021. [Online]. Available: https://software.intel.com/content/dam/develop/external/us/en/documents-tps/multi-key-total-memory-encryption-spec.pdf

[30] AMD. (2016). *AMD Memory Encryption*. Accessed: Apr. 2021. [Online]. Available: https://developer.amd.com/wordpress/media/2013/12/AMD_Memory_Encryption_Whitepaper_v7-Public.pdf

[31] N. Rathi, S. Ghosh, A. Iyengar, and H. Naeimi, "Data privacy in non-volatile cache: Challenges, attack models and solutions," in *Proc. 21st Asia South Pacific Design Automat. Conf. (ASP-DAC)*, Jan. 2016, pp. 348–353.

[32] S. Motaman, S. Ghosh, and N. Rathi, "Cache bypassing and checkpointing to circumvent data security attacks on STTRAM," *IEEE Trans. Emerg. Topics Comput.*, vol. 7, no. 2, pp. 262–270, Apr. 2019.