Literature Review

September 2025

Ian Barnaby

ACM Reference Format:

Ian Barnaby. 2025. Literature Review September 2025. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

This project aims to research and design an asymmetric encryption method targeting a processor and AI accelerator pair. In the theorized application, the accelerator is taken as a trusted base holding the IP of a neural network. It is therefore desired to have an encryption scheme between CPU and accelerator to prevent malicious actors from performing snooping attacks on the bus connecting the accelerator and processor. This project specifically considers accelerators utilizing non-volatile memory to perform compute in-memory operations. These types of devices have been demonstrated to allow for information to be encoded in physical device properties, without being stored in the current state of the cell. In the proposed encryption mechanism, values encoded in the memory cells are used to generate a private/public key pair, used to encrypt processor/accelerator communication.

2 Relevant Background

This literature review focuses on literature regarding the relevant NVM devices/device properties, asymmetric encryption, PUF-based key generation, fault injection defense

3 Search Terms

In searching for papers, databases such as EngineeringVillage and IEEEXplore were used. Search terms included terms such as, but not limited to:

-

4 Paper Reviews

4.1 Hiding Information for Secure and Covert Data Storage in Commercial ReRAM Chips

This paper demonstrates a technique for covert data storage in ReRAM devices [1]. The researchers state that the set/reset time of a ReRAM cell changes with each write cycle. This occurs due to oxygen vacancies in the oxide layer used to construct the conductive filament and limits the total number of write cycles the cell can handle before it no longer functions. The researchers have observed that, once the set/reset characteristics of a cell have been altered, they can reliable determine cells with more or less "wear". By setting a point at which set/reset times can be read as binary "1" or "0", cells can intentially be worn down to encode values in the set/reset time. The researchers demonstrate that fresh cells can be worn down in desired bit patterns, and the data can be be reliably recovered until a threshold is reached. After a certain number of write cycles, the data encoded in the set/reset time becomes too noisy. As such, this paper provides the basis for this work of establishing a method of encoding data within ReRAM cells separate from the memory values, that changes based on how many writes have been issued to the device.

4.2 Overview of NVM Devices

The above paper has proven the potential for data encoding in ReRAM, but other NVM technologies are worthy of review. As NVM and CIM are unsolved fields, commercial applications may move away from ReRAM. There are three other types of NVM technologies deserving particular attention: ECM, phase-change, and magnetic memory. While other technologies exist, they are substantially less relevant or developed for deep learning applications.

4.2.1 ECM. ECM (electrochemical metallization) memory operates on a similar principle to ReRAM: a low resistance state is created by forming a conductive filament between two electrodes, and a high resistance state is created by destruction of the filament. In ReRAM (also called valence change memory), this conductive filament is formed by oxygen vacancies in an exotic metal oxide layer. In ECM, this conductive filament is formed by the migration of metal ions from an active electrode towards an inert electrode through an oxide layer. An existing work demonstrates the set/reset characteristics of ECM in regards to endurance [2], supporting that set/reset characteristics change over many write cycles. While further study would be required to determine if the same technique can be reliably applied to ECM, the required behavior is present.

4.2.2 PCM. PCM (phase change memory) differs in that it does not utilize a conductive filament. In PCM, a layer of thin chalcogenide is deposited between two conducting contacts. By applying a moderate current pulse the chalcogenide layer is heated, forming a conducting crystalline layer to encode a low resistance state. By applying a short, high current pulse the chalcogenide layer is melted, rapidly cooling into a non-conducting amorphous layer to encode a high resistance state. An existing work demonstrates that the set/reset characteristics change over time [4] for determining device endurance. Again, further study would be required to determine if the same technique can be applied, but the required behavior is present.

4.2.3 STT-MRAM. STT-MRAM (spin transfer torque magnetic RAM) operates on a substantially different principle than any of the above technologies. An STT-MRAM cell consists of a reference

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

magnetic layer (magnetization direction is fixed), free magnetic layer (magnetization direction is variable) and a tunneling barrier between the two. The state encoding is determined by the direction of an applied spin-polarized current. When current is applied in the 'forward' direction, spin-polarized electrons align the free layer to the reference layer to encode a low resistance state. When current is applied in the 'backward' direction, the free layer is forced to align opposite to the reference layer, encoding a high resistance state. Existing works studying the endurance of STT-MRAM in deep learning applications demonstrate devices capable of $>10^{12}$ write cycles with little to no timing drift [8]. As MRAM does not rely on a filament creation or phase change, only magnetization direction, this makes intuitive sense. Unlike the previous technologies, STT-MRAM does not display the required behavior, and the encoding technique is unlikely to work.

4.3 Fault Injection Attacks in DNNs

Though this work's main concern is with the protection of IP, it is important to note that deep neural networks (DNNs) have been shown to be highly susceptible to bit flips and fault injection attacks as discussed in [6]. This is especially the case with an edge-AI application where quantized weight data is likely to be used. In these applications one targeted bit flip can cause worse than random guess inference from the model [3]. Our threat model does not consider RowHammer type of exploits(this is further explained in section 4.7), but we consider the bus connection between an AI accelerator and processor to be vulnerable to tampering.

One proposed solution for a fault injection attacks [3] uses low overhead Pearson hashing to create ground truth hashes per DNN level. The hashes are generated from an assumed benign DNN prior to deployment. During execution time the hashes are recomputed per-level parallel and verified against the ground truth hashes at model checkpoints. For mitigations like this one, which rely on the integrity of pre-deployment data, the use of PUF based encryption may be an integral addition to the secure operation of the model by ensuring verification secrets are kept private from advisories over an untrusted line of communication.

4.4 Bus Encryption for Low-Power Systems

This paper, *Sealer: In-SRAM AES for High-Performance and Low-Overhead Memory Encryption* [9], presents a lightweight approach to securing memory and bus transactions in low-power systems. The authors note that conventional bus and memory encryption schemes often rely on AES engines placed in the memory controller, which introduce significant latency and energy overhead since encryption/decryption lies on the critical path of every memory access. To address this, the Sealer architecture repurposes SRAM subarrays to perform AES operations directly within the memory array, exploiting intrinsic bitline-level parallelism. This allows data to be encrypted before leaving the chip and decrypted upon return, thereby protecting against bus snooping and cold-boot attacks without the heavy cost of dedicated crypto hardware. The results demonstrate up to two orders of magnitude improvement in throughput-per-area and a 3× reduction in energy compared to prior solutions. For this project, the relevance lies in the demonstration that bus encryption can be achieved with minimal hardware overhead, suggesting that similar lightweight encryption approaches could be adapted for CPU-to-accelerator communication.

4.5 PUF-Based Encryption Using ReRAM Devices

The paper *An Error Correction Approach to Memristors PUF-based Key Encapsulation* [5] explores the use of ReRAM-based Physically Unclonable Functions (PUFs) for secure key generation and encryption. The authors propose a keyless encapsulation protocol that leverages the inherent resistance variability of ReRAM cells to generate unique device responses, which are then used directly for message encryption. Unlike traditional schemes, this avoids the need for stored cryptographic keys, reducing vulnerability to key extraction attacks. However, because ReRAM PUF responses are sensitive to environmental conditions and device noise, the paper introduces error correction coding (ECC) mechanisms (Reed-Solomon and BCH codes) to stabilize the responses and ensure reliable decryption. Experimental results confirm that the scheme can produce noise-free messages under typical operating variations. For this project, the significance is clear: ReRAM-based PUFs provide a natural hardware root of trust that can generate evolving keys tied to device physics, aligning directly with the proposed idea of leveraging NVM wear characteristics for asymmetric encryption in IMC accelerators.

4.6 Rowhammer Attacks

Rowhammer-type attacks have been studied in certain non-volatile memories and bit flips have been demonstrated [7], but have not been tested in accelerator systems. They will not be considered for this project as proper testing requires hardware and cannot be done in simulation. If rowhammer attacks are shown to be feasible against accelerator systems, a defense would be an important future direction.

5 Summary of Field

References

[1] Farah Ferdaus, B. M. S. Bahar Talukder, and Md. Tauhidur Rahman. 2024. Hiding Information for Secure and Covert Data Storage in Commercial ReRAM Chips. *IEEE Transactions on Information Forensics and Security* 19 (2024), 3608–3619. doi:10.1109/TIFS.2024.3364845

[2] J. Q. Huang, L. P. Shi, E. G. Yeo, K. J. Yi, and R. Zhao. 2012. Electrochemical Metallization Resistive Memory Devices Using ZnS-SiO_2 as a Solid Electrolyte. *IEEE Electron Device Letters* 33, 1 (2012), 98–100. doi:10.1109/LED.2011.2173457

[3] Mojan Javaheripi and Farinaz Koushanfar. 2021. HASHTAG: Hash Signatures for Online Detection of Fault-Injection Attacks on Deep Neural Networks. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. 1–9. doi:10.1109/ICCAD51958.2021.9643556

[4] Wu Lei, Cai Daolin, Chen Yifeng, Liu Yuanguang, Yan Shuai, Li Yang, Yu Li, Xie Li, and Song Zhitang. 2021//. Impact of Continuous RESET/SET Operations on Endurance Characteristic of Phase Change Memory. *Journal of Shanghai Jiao Tong University* 55, 9 (2021//), 1134 – 41.

[5] M. Liu, B. Yan, et al. 2021. An Error Correction Approach to Memristors PUF-based Key Encapsulation. In *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*. 1–6. doi:10.1109/COINS51742.2021.9524282

[6] Hafizur Rahaman, Chandan Giri, Surajit K Roy, and Amlan Chakrabarti. 2025. Optimization and Security of AI Models for Deployment at Edge: A Comprehensive Review. In *2025 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Vol. 1. 1–6. doi:10.1109/ISVLSI65124.2025.11130213

[7] Felix Staudigl, Hazem Al Indari, Daniel Schön, Dominik Sisejkovic, Farhad Merchant, Jan Moritz Joseph, Vikas Rana, Stephan Menzel, and Rainer Leupers. 2022. NeuroHammer: Inducing Bit-Flips in Memristive Crossbar Memories. In *2022 Design, Automation and Test in Europe Conference and Exhibition (DATE)*. 1181–1184. doi:10.23919/DATE54114.2022.9774651

[8] Z. Wei, W. Kim, Z. Wang, L. Hu, D. Jung, J. Zhang, and Y. Huai. 2022. Accurate and Fast STT-MRAM Endurance Evaluation Using a Novel Metric for Asymmetric Bipolar Stress and Deep Learning. In *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. 373–374. doi:10.1109/VLSITechnologyandCir46769.2022.9830351

[9] J. Zhang, H. Naghibijouybari, and E. Sadredini. 2022. Sealer: In-SRAM AES for High-Performance and Low-Overhead Memory Encryption. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED '22)*. 1–6. doi:10.1145/3531437.3539699