D22-037

具安全防護之自旋式三合一記憶體內運算應用於小型 AI 邊緣運算裝置

CMOS-integrated Security-Aware Spintronics Three-in-one Compute-in-Memory Macro for Tiny AI Edge Devices

隊伍名稱　iMDL 堅強陣容
iMDL Tough Team

隊　　長　邱硯晟 / 清華大學電機工程研究所
隊　　員　謝鈁淩 / 清華大學電子工程研究所
簡佑安 / 清華大學電子工程研究所
李忠遠 / 清華大學電機工程研究所

指導教授

張孟凡　清華大學電機工程學系

交通大學電子工程博士、美國賓州州立大學碩士，現為清華大學電機工程學系特聘教授，曾任科技部微電子學門召集人、IEEE 中華民國分會理事長，並具有工業界工作 13 年經驗。

研究領域

記憶體內運算電路設計、記憶體之安全與區塊鍊電路設計、自旋電路與非揮發邏輯電路設計

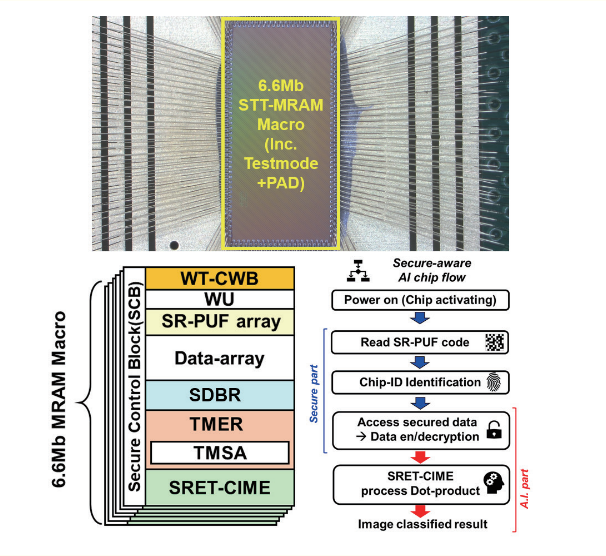
作品摘要

電池供電的AI邊緣運算裝置需要高準確率、高效率運算及具耐受度高和高能效電池等特性，再加上日趨重要的硬體安全，記憶體內資料存儲和讀取需嚴加保護，因此提出第一個具使用安全認證、在斷電時保護儲存在非揮發性記憶體資料和非揮發性記憶體內運算（nvCIM）之三合一功能之電路巨集並應用於AI邊緣運算的晶片，不但能以高能效完成高效率運算，也運用金鑰驗證確保晶片的資料在供電和斷電下都不被竊取。

我們提出世界第一個擁有安全考量功能的非揮發性三合一記憶體內運算巨集。此巨集為22奈米製程技術，且具有6.6Mb自旋力矩轉移磁阻式隨機存取記憶體（STT-MRAM）。此巨集可支援（1）晶片啟動時的身分認證、（2）存取加密資料、（3）高能源效率人工智慧相關記憶體內運算。為了達成上述三合一功能之巨集，我們提出並實作出以下想法：

1. 我們使用STT-MRAM寫入時間的不同來做為亂度，提出基於自旋電子特性之可重構物理不可克隆函數（SR-PUF），並用SR-PUF產生的輸出來當作各個晶片獨特的晶片指紋，搭配雜湊函數（SHA-256）來實現晶片啟動時的使用安全認證。
2. 我們使用正反雙向加密特性（2DHC-PE）來加密存取於非揮發性記憶體的資料，並利用自我解密的突發讀取方式（SDBR）來快速解密受到加密保護的資料。
3. 我們利用sparsity和ReLU打造一個能夠感知稀疏性和ReLU函數的提前終止之記憶體內運算引擎（SRET-CIME）來減少運算量並達到高能源效率之目且不會影響運算準確度。此巨集最高可以支援到八位元輸入、八位元權重、五百七十六累加數之乘加運算。

我們的晶片是使用晶圓廠提供的22奈米製程和STT-MRAM。總共的晶片面積是18mm^2（6mm*3mm）。有3456個巨集輸出數量且讀取頻寬可達到288GB/s，在目前所有非揮發性記憶體作品中，具有最高的讀取頻寬。所提出的SR-PUF具有inter-HD:0.4999和intra-HD:0。本作品最高可支援八位元輸入、八位元權重和二十六位元輸出，能量效率為30.1~68TOPS/W。也為目前nvCIM作品中最高。



圖一　STT-MRAM nvCIM 巨集晶片照、架構及流程圖

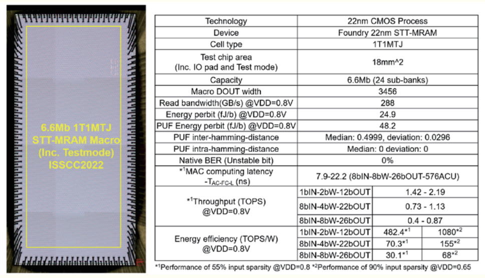
Abstract

AI edge devices with high inference accuracy, rapid response times, and long battery life require high energy efficiency. Ensuring the security of devices against malicious attack or illegal access also requires data protection mechanisms and secure access control. This work presents the first ever three-in-one security-aware nonvolatile compute-in-memory (nvCIM) macro for edge AI chips, featuring secure access control, data protection against power-on and power-off probing , and high energy-efficiency computing capability.

This work fabricated a 6.6Mb CMOS-integrated security-aware three-in-one nvCIM macro using foundry-provided 22nm spin-transfer torque magnetic random-access memory technology. The macro can support (1) chip-ID identification during chip activating (2) encrypted data storage (3) high-energy efficiency AI-related computing-in-memory operation. Primary contributions of this security-aware three-in-one nvCIM macro are listed as follows:

1. We developed a spintronic-based reconfigurable physically unclonable function (SR-PUF) with high randomness and high reliability based on the difference of STT-MRAM write time. The outcome of SR-PUF can be taken as a unique chip fingerprint then coupled with hash function to realize chip-ID identification when chip activating.
2. We developed a 2-dimensional half-complement physical encryption (2DHC-PE) scheme and corresponding snoop-proof self-decryption burst-read (SDBR) scheme, enabling access to encrypted data with low read latency and uniform power noise.
3. We developed a sparsity-and-ReLU-aware early-termination computing-in-memory engine (SRET-CIME) to reduce the number of partial dot-product operations in order to reduce energy efficiency without degrading system-level accuracy. This scheme has a 26b output to support full-channel dot-product operations of up to 8b-input and 8b-weight with 576 accumulations.

The proposed 6.6Mb nvCIM macro was fabricated under 22nm technology and using foundry-provided 22nm spin-transfer torque magnetic random-access memory. Chip area is 18mm^2 (6mm*3mm). The macro has 3456 macro DOUT width and a read bandwidth of 288GB/s. The macro also achieved high randomness (inter-hamming-distance: 0.4999) and high reliability (intra-hamming-distance: 0) for PUF functions with high energy-efficiency (30.1~68TOPS/W) and short latency (7.9~22.2ns) for high-precision dot-product operations (8b-input, 8b-weight, 576 accumulation, and 26b output). These three specifications are outperformed all published nvCIM works.


6.6Mb 1T1MTJ
STT-MRAM Macro
(Inc. Testmode)
ISSCC2022

Technology	22nm CMOS Process		
Device	Foundry 22nm STT-MRAM		
Cell type	1T1MTJ		
Test chip area (Inc. IO pad and Test mode)	18mm^2		
Capacity	6.6Mb (24 sub-banks)		
Macro DOUT width	3456		
Read bandwidth(GB/s) @VDD=0.8V	288		
Energy perbit (fJ/b) @VDD=0.8V	24.9		
PUF Energy perbit (fJ/b) @VDD=0.8V	48.2		
PUF inter-hamming-distance	Median: 0.4999, deviation: 0.0296		
PUF intra-hamming-distance	Median: 0 deviation: 0		
Native BER (Unstable bit)	0%		
*1MAC computing latency -$T_{AC-FC-L}$ (ns)	7.9-22.2 (8bIN-8bW-26bOUT-576ACU)		
*1Throughput (TOPS) @VDD=0.8V	1bIN-2bW-12bOUT	1.42 - 2.19	
	8bIN-4bW-22bOUT	0.73 - 1.13	
	8bIN-8bW-26bOUT	0.4 - 0.87	
Energy efficiency (TOPS/W) @VDD=0.8V	1bIN-2bW-12bOUT	482.4*1	1080*2
	8bIN-4bW-22bOUT	70.3*1	155*2
	8bIN-8bW-26bOUT	30.1*1	68*2

*1Performance of 55% input sparsity @VDD=0.8 *2Performance of 90% input sparsity @VDD=0.65

Fig. 2 Chip summary