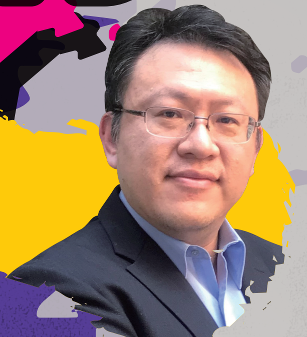
D23-056

適用於非揮發AI邊緣處理器具有4MB的單級-多級單元混合型可變電阻式和靜態隨機存取記憶體內運算晶片

A Nonvolatile AI-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM-SRAM Compute-in-Memory-Fusion Unit

隊伍名稱｜Tapeout不揪
Tapeout Alone
隊　　長｜溫戴豪 / 清華大學電機工程研究所
隊　　員｜洪哲民 / 清華大學電機工程研究所
張傳佳 / 清華大學電機工程研究所
許宏禧 / 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。

研究領域

記憶體內運算電路設計（深度學習人工智慧晶片應用）、記憶體積體電路設計、使用記憶體之安全與區塊鏈電路設計、仿生人工智慧晶片之憶阻器電路設計、低功率及低電壓積體電路設計、自旋電路與非揮發邏輯電路設計、新奈米元件之電路設計

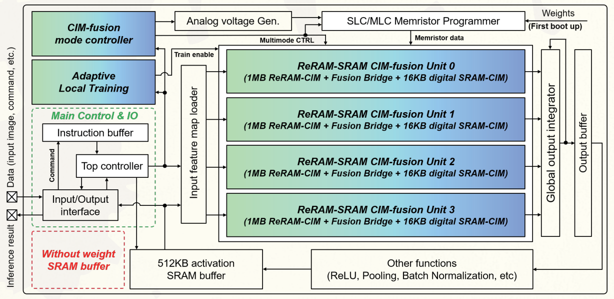
作品摘要

人工智能（AI）邊緣設備更傾向於採用高容量的非揮發性記憶體內運算（nvCIM）來實現高能源效率和快速的喚醒至回應時間，同時保持足夠的推理準確性。以前的多級單元（MLC）ReRAM-CIM，因為與製造相關的變異性，在CMOS相容性和推理準確性方面存在局限，現有的商用CMOS相容ReRAM設備僅保證單級單元（SLC）行為，並且由於耐久性有限，不支持本地訓練。

現有的靜態隨機存取記憶體內運算（SRAM CIM）及數位訊號處理器（Digital Processor），在喚醒之後，資料需要從外部非揮發性記憶體移至內部暫存器和運算單元，耗費大量能量和時間，而非揮發性記憶體內運算則藉由儲存和運算在同晶片上，抑制延遲與能耗。

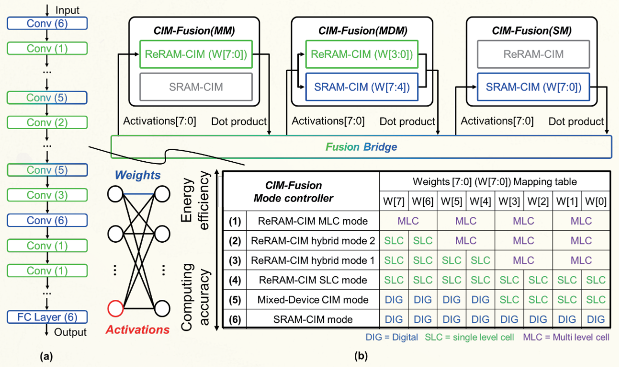
我們提出使用了ReRAM-SRAM fusion為基準的22奈米人工神經網路處理器（圖一），運用全新的架構，完成高效能、低延遲、高精度，並能過支援本地訓練的記憶體內運算晶片處理器。為了突破記憶體內運算的挑戰，我們提出：（1）首個使用商用ReRAM與ReRAM-SRAM CIM-fusion方案結合的CMOS集成AI邊緣處理器（圖二）；（2）支援本地訓練，應變製造相關的變異性；（3）可調式儲存單元（SLC、SLC-MLC和MLC模式）以平衡面積和讀取良率；（4）動態累加電流量化（DACQ）以抑制位元線上的波動和ReRAM記憶體能量消耗；（5）電流電壓混合類比數位轉換器（CVH-ADC）增加訊號邊距和良率。

本作品是第一款完全集成CMOS的ReRAM-SRAM CIM-fusion人工智能邊緣處理器（圖三），以克服在功能、性能和可製造性方面的個別挑戰。該處理器使用ReRAM-SRAM CIM-fusion結構、軟硬體協同優化和晶片上自適應本地訓練，多級單元（MLC）提供了高儲存密度，具有足夠的記憶體容量來存儲整個神經網絡的參數。該處理器在各種神經網路模型和數據集之間實現了推理準確性和能量效率之間的各種平衡取捨實現具有接近MLC的高容量、短喚醒至回應延遲時間（392.46微秒）、高峰值能源效率（71.2 TOPS/W）和穩健的準確性，並且對於製造差異性方面支持本地訓練。



CIM-fusion mode controller
Adaptive Local Training
Main Control & IO
Instruction buffer
Command
Top controller
Input/Output interface
Data (input image, command, etc.)
Inference result
Without weight SRAM buffer
Analog voltage Gen.
SLC/MLC Memristor Programmer
Weights (First boot up)
Multimode CTRL
Memristor data
Train enable
Input feature map loader
ReRAM-SRAM CIM-fusion Unit 0 (1MB ReRAM-CIM + Fusion Bridge + 16KB digital SRAM-CIM)
ReRAM-SRAM CIM-fusion Unit 1 (1MB ReRAM-CIM + Fusion Bridge + 16KB digital SRAM-CIM)
ReRAM-SRAM CIM-fusion Unit 2 (1MB ReRAM-CIM + Fusion Bridge + 16KB digital SRAM-CIM)
ReRAM-SRAM CIM-fusion Unit 3 (1MB ReRAM-CIM + Fusion Bridge + 16KB digital SRAM-CIM)
Global output integrator
Output buffer
512KB activation SRAM buffer
Other functions (ReLU, Pooling, Batch Normalization, etc)

圖一 ReRAM-SRAM fusion的神經網路處理器。



Input
Conv (6)
Conv (1)
Conv (5)
Conv (2)
Conv (5)
Conv (3)
Conv (6)
Conv (1)
Conv (1)
FC Layer (6)
Output
Weights
Activations
(a)
CIM-Fusion(MM)
ReRAM-CIM (W[7:0])
SRAM-CIM
CIM-Fusion(MDM)
ReRAM-CIM (W[3:0])
SRAM-CIM (W[7:4])
CIM-Fusion(SM)
ReRAM-CIM
SRAM-CIM (W[7:0])
Activations[7:0]
Dot product
Fusion Bridge
Energy efficiency
Computing accuracy
(b)

CIM-Fusion Mode controller		Weights [7:0] (W[7:0]) Mapping table							
		W[7]	W[6]	W[5]	W[4]	W[3]	W[2]	W[1]	W[0]
(1)	ReRAM-CIM MLC mode	MLC		MLC		MLC		MLC	
(2)	ReRAM-CIM hybrid mode 2	SLC	SLC	MLC		MLC		MLC	
(3)	ReRAM-CIM hybrid mode 1	SLC	SLC	SLC	SLC	MLC		MLC	
(4)	ReRAM-CIM SLC mode	SLC	SLC	SLC	SLC	SLC	SLC	SLC	SLC
(5)	Mixed-Device CIM mode	DIG	DIG	DIG	DIG	SLC	SLC	SLC	SLC
(6)	SRAM-CIM mode	DIG	DIG	DIG	DIG	DIG	DIG	DIG	DIG

DIG = Digital　SLC = single level cell　MLC = Multi level cell

圖二 CIM-fusion模式控制器與可調式儲存單元模式。

Abstract

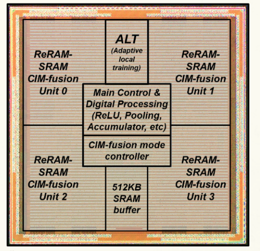
Artificial Intelligence (AI) edge devices tend to adopt high-capacity non-volatile memory for compute-in-memory (CIM) to achieve high energy efficiency and fast wake-up-to-response time while maintaining sufficient inference accuracy. Previous Multi-Level Cell (MLC) ReRAM-CIM solutions had limitations in terms of CMOS compatibility and inference accuracy due to process-related variations. Existing commercial CMOS-compatible ReRAM devices only guarantee Single-Level Cell (SLC) behavior and do not support on-chip local training due to limited endurance.

The existing Static Random Access Memory Computing-In-Memory (SRAM CIM) and Digital Processor require data to be transferred from external non-volatile memory to internal registers and computational units after waking up, consuming a significant amount of energy and time. In contrast, non-volatile CIM performs storage and computation on the same chip, suppressing latency and power consumption.

We propose a 22-nanometer artificial neural network processor (Figure 1) based on the ReRAM-SRAM fusion architecture, which achieves high performance, low latency, high precision, and supports on-chip local training. To overcome the challenges of in-memory computation and other processor-level challenges, we proposed: (1) the first CMOS-integrated AI edge processor that combines commercial ReRAMs with ReRAM-SRAM CIM-fusion architecture (Figure 2); (2) support for on-chip local training to adapt to process-related variations; (3) adjustable storage units (SLC, SLC-MLC, and MLC modes) to balance area and yield; (4) Dynamic Accumulator Current Quantization (DACQ) to suppress fluctuations on bitlines and energy consumption of variable-resistance memories; (5) current-voltage hybrid analog-to-digital converter (CVH-ADC) to increase signal margins and yield.

This work presents the first fully CMOS-integrated ReRAM-SRAM CIM-fusion AI edge processor (Figure 3) to overcome individual challenges in terms of functionality, performance, and manufacturability. The processor utilizes the ReRAM-SRAM CIM-fusion structure, software-hardware co-optimization, and on-chip adaptive local training. The MLC ReRAM cell provides high storage density with sufficient memory capacity to store the parameters of the entire neural network. The proposed processor achieves various trade-offs between inference accuracy and energy efficiency across a variety of neural network models and datasets. It provided high CIM capacity due to MLC, short wake-up-to-response latency (392.46 μs), high peak energy efficiency (71.2 TOPS/W), and robust accuracy while supporting on-chip local training for manufacturing variations.

4MB CIM-fusion AI edge processor



ReRAM-SRAM CIM-fusion Unit 0
ALT (Adaptive local training)
ReRAM-SRAM CIM-fusion Unit 1
Main Control & Digital Processing (ReLU, Pooling, Accumulator, etc)
CIM-fusion mode controller
ReRAM-SRAM CIM-fusion Unit 2
512KB SRAM buffer
ReRAM-SRAM CIM-fusion Unit 3

Technology	22nm CMOS logic process (Ultra low leakage)	
ReRAM	Foundry provided 1T1R ReRAM	
Test chip area	30.6mm² (Include IO pad, testmodes)	
Application	CNN/FC	
Computing approach	ReRAM-CIM, Mixed-Device CIM, SRAM-CIM	
Frequency	50 - 200MHz	
ReRAM memory	4M Byte	
SRAM buffer	512K Byte	
Supply voltage	0.7 - 0.8V	
Input / Wight precision	1b to 8b (INT)	
Throughputs*4 (TOPS)	INT 8	6.97 *1
Energy efficiency*4 (TOPS/W)	INT 8	32.2 *2 – 71.2 *3
Computing density*4 (TOPS/mm²)	INT 8	0.284 *1

*1 Peak throughput measured under 0.8V, 200 MHz.
*2 Measured under 0.8V, 200 MHz using MobileNet-v2 trained for ImageNet.
*3 Measured under 0.7V, 50 MHz using ResNet-20 trained for MNIST.
*4 One operation (OP) represents one multiplication or one addition.

Fig.3 Die photo, CIM-fusion controller, and adaptive local training.