D21-001

適用於 AI 邊緣設備具有 11.91 到 195.7 TOPS/W 的 22 奈米 4Mb 8 位元精度可變電阻式記憶體內運算晶片

A 22nm 4Mb 8b-Precision ReRAM Computing-in-Memory Macro with 11.91 to 195.7 TOPS/W for Tiny AI Edge Devices

隊伍名稱
下線好趕
Tape-out So Hurry
隊長
洪哲民
清華大學電機工程研究所

隊員
張富淳
清華大學電子工程研究所
温戴豪
清華大學電機工程研究所
黃彥翔
清華大學電機工程研究所

作品摘要

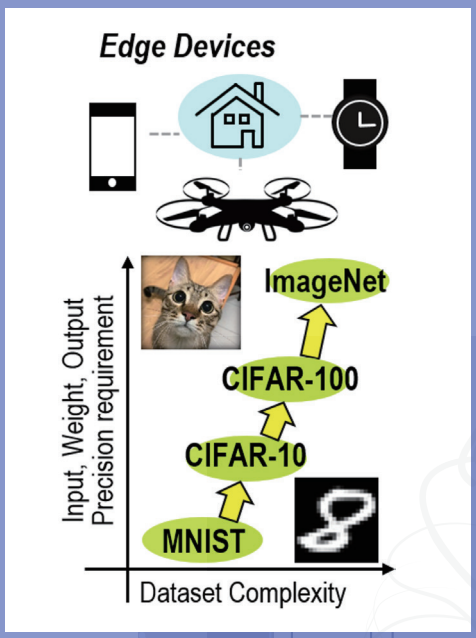
隨著深度學習和物聯網的發展，需要的精度與計算資料量隨著神經網路的複雜度而上升（圖一），電池供電的AI邊緣裝置為了支持更複雜的運用，需要具備大記憶體容量（capacity）、多位元輸入（IN）、權重（W）與輸出（OUT）精度的非揮發性記憶體內運算（nvCIM）晶片來實現高能效（EFMAC）、低延遲（tAC）的乘加（MAC）運算，目前受限於低的讀取電壓上限與位元線（BL）上過大的寄生電容，大多數非揮發性記憶體內運算（nvCIM）使用電流讀取架構，且僅能達到低的輸入和權重精度運算（binary to 4b），本作品期望能運用全新的架構，來解決此問題。

我們提出了使用SLC 1T1R可變電阻式記憶體技術的22奈米4Mb可變電阻式記憶體巨集（圖二），巨集的乘加運算由通道方向（bitline）上的4組8位元輸入8位元權重組成，8位元權重以2 補數（2's complement）格式儲存在同一列（WL）的8個可變電阻式記憶體單元中，每一行的組成為記憶體單元、行的數據多工器（column MUX）、具權重的電流轉電壓訊號疊加（WCVSS）轉換器、電壓型感測放大器（VSA）和數位位移與加法電路（DSAC）。

為了突破記憶體內運算的挑戰，我們提出以下作品特色: 1. 非對稱輸入群組調變架構（AGMI），切分8位元輸入至子群組（2b-3b-3b）以降低運算延遲，並讓最高有效位（MSBs）保有足夠的訊號邊距（signal margin）; 2. 具權重的電流轉電壓訊號疊加（WCVSS）轉換器，轉換部分乘加的資料線電流（IDL）為電壓的訊號，並保有其位置權重; 3. 運用電壓型感測放大器（VSA）的混和精度電壓型讀取架構，以減低能量消耗和降低多位元乘加的的讀取延遲（tAC），並仍讓最高有效位有足夠的訊號邊距。

本作品完成了具有11.91到195.7 TOPS/W的22奈米 4Mb 8位元精度可變電阻式記憶體內運算晶片，是第一片支持8位元輸入8位元權重乘加的記憶體內運算巨集，支援從單位元輸入與權重到8位元輸入與權重的乘加，並在目前經過矽驗證的非揮發記憶體內運算中，達到最快的計算延遲與最佳的能耗。

圖三顯示了裸晶照片及匯總表。



▲ 圖一 邊緣裝置要求的精度隨著複雜應用而增加

指導教授

張孟凡 清華大學電機工程學系

交通大學電子博士，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

奈米及記憶體電路設計、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、記憶體內運算電路設計、人工智慧晶片與深度學習應用之憶阻器電路設計

Abstract

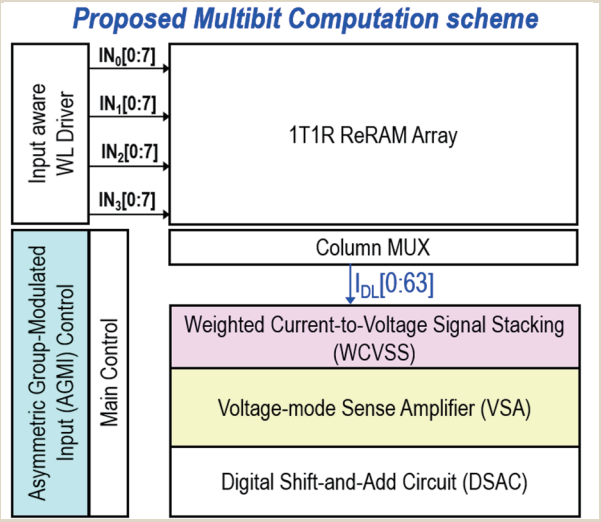
Battery-powered tiny-AI edge devices require large-capacity nonvolatile compute-in-memory (nvCIM), with multibit input (IN), weight (W), and output (OUT) precision to support complex applications (Fig.1), high-energy efficiency (EFMAC), and short computing latency (tAC) for multiply-and-accumulate (MAC) operations. Due to the low read-disturb-free voltage of nonvolatile memory (NVM) devices and the large parasitic load on the bitline, most existing Mb-level nvCIM macros use a current-mode read scheme and only achieve a low IN-W precision (binary to 4b).

Our nvCIM macro delivers MAC operations comprising 4 sets of 8bIN and 8bW in the channel direction (Fig.2). 8bW data is stored in 8 ReRAM cells on the same row (wordline/WL) across 8 columns in 2's complement format. Each set of columns comprises memory cells, column multiplexors (column MUX), a WCVSS converter, a VSA, and digital shift-and-add circuits (DSAC).

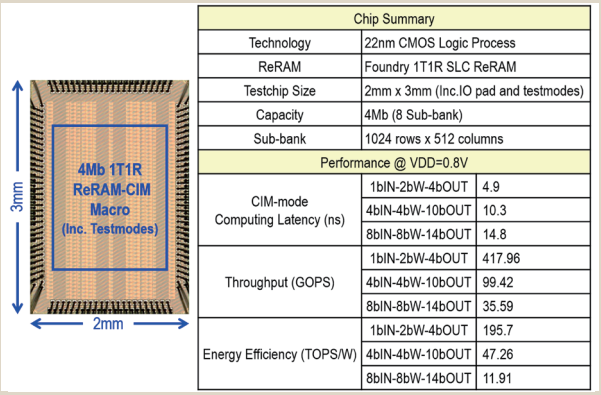
To overcome these challenges we developed: 1. An asymmetric group-modulated input (AGMI) scheme, in which a 8b-input is split into 3 sub-groups (2b-3b-3b) to reduce computing latency, while maintaining sufficient signal margins for the most significant bits (MSBs). 2. A weighted current-to-voltage signal stacking (WCVSS) converter to translate the dataline current (IDL) of partial MAC operations into voltage-mode signals with respective place values. 3. A hybrid-precision voltage-mode readout scheme with voltage-mode sense amplifier (VSA) to reduce energy consumption and shorten the tAC for multibit MAC readout operations, while maintaining the sensing margin for MSBs.

The proposed 22nm 4Mb ReRAM macro, fabricated using foundry SLC 1T1R ReRAM technology, is the first nvCIM macro supporting 8b-input and 8b-weight MAC operations. Among silicon-verified nvCIMs, it achieved the fastest tAC (4.9-14.8ns) and best EFMAC (195.7-11.91TOPS/W) with precision from binary IN-W to 8bIN-8bW-14bOUT.

Figure 3 shows a die photo and summary table.



▲ Fig. 2 Proposed multibit computation scheme



Chip Summary		
Technology	22nm CMOS Logic Process	
ReRAM	Foundry 1T1R SLC ReRAM	
Testchip Size	2mm x 3mm (Inc.IO pad and testmodes)	
Capacity	4Mb (8 Sub-bank)	
Sub-bank	1024 rows x 512 columns	
Performance @ VDD=0.8V		
CIM-mode Computing Latency (ns)	1bIN-2bW-4bOUT	4.9
	4bIN-4bW-10bOUT	10.3
	8bIN-8bW-14bOUT	14.8
Throughput (GOPS)	1bIN-2bW-4bOUT	417.96
	4bIN-4bW-10bOUT	99.42
	8bIN-8bW-14bOUT	35.59
Energy Efficiency (TOPS/W)	1bIN-2bW-4bOUT	195.7
	4bIN-4bW-10bOUT	47.26
	8bIN-8bW-14bOUT	11.91

▲ Fig. 3 Die photo and chip summary