D19-080

應用於深度學習 AI 邊緣運算處理器之 14.6 奈秒平行運算讀取速度 1Mb 電阻式記憶體內運算巨集

A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN-Based AI Edge Processors

隊伍名稱 阿瑞捲雞
Ares Chicken Roll

隊　　長 薛承昕 / 清華大學電子工程研究所

隊　　員 劉哲旭 / 清華大學電機工程研究所
魏瑋辰 / 清華大學電機工程研究所
魏士穎 / 清華大學電子工程研究所

DESIGN GROUP D19-080

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，現為清華大學電機工程學系教授。曾任積丞科技矽智產事業處處長、台積電設計服務處主任工程師、美國 Mentor Graphics 工程師。

研究領域
記憶體積體電路設計、非揮發性邏輯電路設計、人工智慧晶片之記憶體內運算電路設計。

指導教授
鄭桂忠
清華大學電機工程學系

美國加州理工學院電機博士，現為清華大學電機工程學系教授。曾任美國 Second Sight Medical Products 公司資深電機工程師。

研究領域
人工智慧晶片、仿神經晶片、生醫訊號處理、仿生系統、生醫系統、微型電子鼻系統、氣體感測、類比及混合信號積體電路設計、生醫電子晶片設計。

作品摘要

傳統處理器架構中，運算資料於處理器與記憶體之間透過傳輸線（Bus）進行傳遞稱馮諾伊曼（Von Neumann）架構。隨著大數據技術與 AI 晶片發展，系統運算的資料量出現突破性的增加，在傳統架構中資料於記憶體與處理器間傳輸介面的輸入輸出端（IO）數成為速度瓶頸，而搬動資料需要消耗大量額外能量亦成為效能上的限制。近年，記憶體內運算成為目前最具潛力的研究項目。有別於傳統馮諾伊曼架構，記憶體內運算可於單晶片中實現平行運算，降低需要傳遞與暫存的資料量達到快速且低功耗之運算目標。

本作品使用利用高密度、高低阻態比值（R-ratio）大的電阻式記憶體（ReRAM）提出創新之非揮發性記憶體內運算巨集（Nonvolatile Computing-In-Memory，nvCIM），並應用於深度學習（Deep Learning，DL）神經網路中進行系統驗證。本作品之記憶體內計算巨集之記憶體不僅可作為存取單元並可於記憶體中進行資料運算，可有效降低資料傳輸量與多餘能量損耗。目標為應用於下世代能量與硬體資源有限之 AI 邊緣裝置（edge device），以下為本作品之電路特色：

1. 國際發表中操作速度最快（11.7奈秒）的非揮發性記憶體內運算巨集。
2. 國際首次，nvCIM 為基底之 CIFAR-10 dataset 系統整合驗證，辨識成功率高達 88.52%。
3. 記憶體內運算電路開發。

本作品提出 Serial-Input Non-Weighted Product（SINWP）structure 可降低做乘加運算之功率消耗。

Down-Scaling Weighted Current Translator（DSWCT）可大幅度的降低加總電流，更能有效的提升讀取良率以及降低功率消耗。

Triple-Margin Current-mode Sense Amplifier（TMCSA）與傳統感測放大電路相比降低6倍的感測電流飄移（Ioffset）。

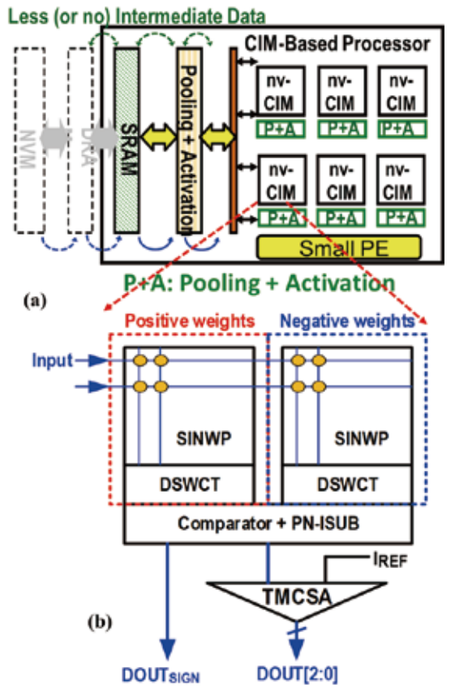

圖 1. (a)搭載多位元非揮發性記憶體內運算巨集之DNN處理器
(b)多位元非揮發性記憶體內運算巨集概覽

ABSTRACT

In the traditional processor architecture, the data is transferred between the processor and the memory through a transmission line (Bus) called the Von Neumann architecture. With the development of big data technology and AI chip, the amount of data in system computing has increased dramatically. In the traditional architecture, the number of input and output of data transmission interface between memory and processor becomes the speed bottleneck. And data movement needs to consume a lot of extra energy and is also a performance limitation.
In recent years, Computing-In-Memory has become the most potential research project. Different from the traditional von Neumann architecture, Computing-In-Memory can achieve parallel operations in a single chip, reducing the amount of data that needs to be transferred and temporarily stored to achieve fast and low power computing goals.

This work uses a high-density, high R-ratio resistive memory (ReRAM) to propose an innovative nonvolatile Computing-In-Memory (nvCIM) and is applied to System verification in Deep Learning (DL) neural networks. The memory of the computational macro in the memory of this work can not only serve as an access unit but also perform data operations in the memory, which can effectively reduce the amount of data transmission and excess energy loss. The target is an AI edge device with limited energy and hardware resources for the next generation. The following are the circuit features of this work:

1. The fastest (11.7 nanoseconds) non-volatile memory computing macro in international publications.
2. For the first time in the world, nvCIM-based CIFAR-10 dataset system integration verification, and the recognition success rate is as high as 88.52%.
3. Computing-In-Memory circuit development.

This work proposes a Serial-Input Non-Weighted Product (SINWP) structure to reduce the power consumption of multiply-add operations.

Down-Scaling Weighted Current Translator (DSWCT) can greatly reduce the total current, and can effectively improve the reading yield and reduce the power consumption.

The Triple-Margin Current-mode Sense Amplifier (TMCSA) reduces the sense current drift (Ioffset) by a factor of 6 compared to conventional sense amplifier circuits.