設計組
Design Group

D18-004

應用於AI晶片之非揮發性記憶體內運算巨集與二進位深度神經網路

Non-volatile Computing-in-Memory Macro Based Binary-Input Ternary-weight Neural Network in Application of AI Chip

隊伍名稱　恩恩阿睿 NNRRAM
隊　　長　陳韋豪 / 清華大學電子工程研究所
隊　　員　李品逸 / 清華大學電機工程研究所
　　林威宇 / 清華大學電子工程研究所
　　薛承昕 / 清華大學電子工程研究所

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。

本研究使用利用高密度、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。本研究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），以下為本研究之電路特色：

1. 國際發表中容量最大且操作速度最快的非揮發性記憶體內運算巨集。應用於下世代AI晶片將可達到1000倍的能量節省。

2. 國際首次，RRAM based nvCIM 為基底之文字辨識（MNIST database）系統整合驗證，系統辨識成功率高達98.8%。

3. 記憶體內運算電路開發。

   - 本研究提出Distance-Racing Current-mode Sense Amplifier（DR-CSA）與傳統感測放大電路相比降低5倍的感測電流飄移（Ioffset）。

   - Input-Aware dynamic IREF（IA-REF）參考電流生成方案提升訊號裕度（Signal margin）由-27.9uA提升至7.8uA。

   - 結合DR-CSA與IA-REF兩種電路應用於DNN文字辨識（MNIST database）中，相較於傳統架構可降低50倍的錯誤發生率。

4. Binary-input Ternary weight network演算法開發。

   - 國際首次，由記憶體內運算電路出發改良之低記憶體需求且兼具高辨識成功率深度神經網路（Deep Neural Network, DNN）模型。

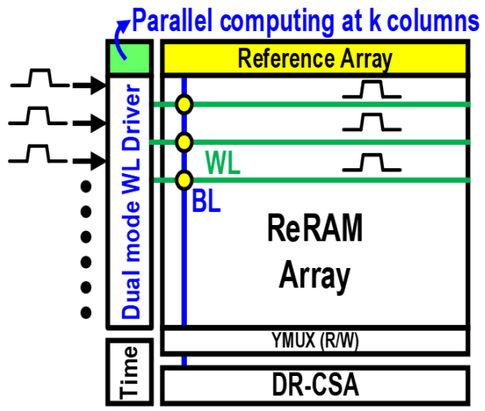
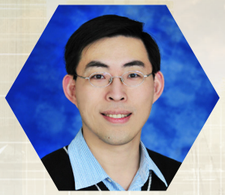


圖1. 利用電阻式記憶體之非揮發性記憶體內運算巨集

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。曾任美國Mentor Graphics工程師、台積電設計服務處主任工程師、積丞科技矽智產事業處處長。

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Second Sight Medical Products, Inc. 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。

Abstract

The challenges faced by von Neumann architecture stem from large amounts of data transmission though memory hierarchies to processing elements (PEs) by bus. Due to the limited IO bandwidth, it not only consumes large energy, but also leads to significant delays. Recently, nonvolatile Computing-in-Memory (nvCIM) becomes a promising solution that enables highly energy-efficient computing for AI edge devices. In particular, nvCIM can achieve fast speed, high throughput and low power consumption by parallel processing.

A 1Mb nvCIM macro was fabricated using 65nm CMOS process with 1T1R contact RRAM (CRRAM) devices. This nvCIM macro can achieve both memory and multiply and accumulation (MAC) CIM functions. The main contributions of this work are listed as follows:

1. The largest capacity and the fastest speed non-volatile computing in memory macro. 1000x energy reduction for the application of AI chip.

2. For the frsit time, record-high 98.8% inference accuracy on MNIST digits recognition has been achieved by nvCIM based demo system.

3. Innovative Computing in memory circuits:

   - 5x input offset reduction by proposed Distance-Racing Current-mode Sense Amplifier (DR-CSA) compare to Converntional Current-mode Sense Amplifier (CNV-CSA).

   - Signal margin improvement from -27.9 uA to 7.8 uA by proposed Input-Aware dynamic Input-Aware dynamic reference generation scheme (IA-REF) reference generation scheme.

   - 50x inference error rate reduction with MNIST database by DR-CSA + IA-REF Scheme.

4. For the first time, hardware driven Binary-input Ternary weight network.

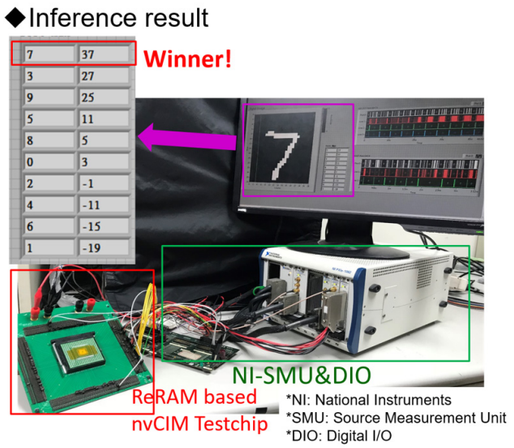


Fig.2 nvCIM based inference system