D21-047

使用電荷運算機制之記憶體內運算之靜態隨機存取記憶體

A Computing-In-Memory SRAM Using Charge-Domain Computing Mechanism

隊伍名稱
騎於山中
Cycling in Mountains
隊長
許尹端
臺灣大學電子工程學研究所

隊員
姚鈞嚴
臺灣大學電子工程學研究所
吳宗諺
臺灣大學電子工程學研究所

作品摘要

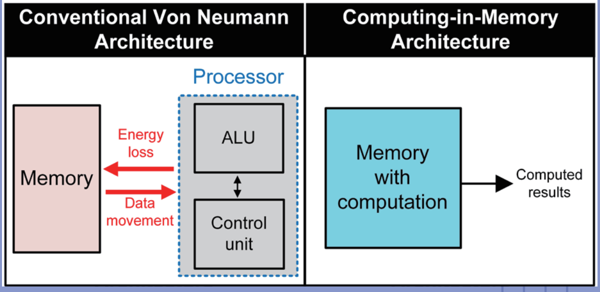
隨著人工智慧（Artificial intelligence，AI）與機器學習（Machine learning，ML）的技術不斷進步，基於神經網路之機器學習架構在語音及影像辨識等應用上已達到出色的準確率。而伴隨著物聯網（Internet of Things，IoT）的發展，將這些機器學習的架構應用於終端設備上面可以獲得諸多好處。相較於傳統的雲端運算，邊緣運算能夠實現較低運算延遲以及較佳的效能。同時，由於不用將資料上傳雲端，避免了資料被第三方竊取的風險，能夠有效提高資料的安全性，也能降低設備對於網絡的依賴性。然而，邊緣運算受限於終端設備之能量與運算資源，使得於終端設備上實現複雜的機器學習架構，變得極具挑戰性。

為因應新興的終端AI應用，記憶體內運算（Computing-in-memory，CIM）之記憶體架構逐漸嶄露頭角。藉由直接在記憶體內進行運算以避免大量的資料搬運，此記憶體架構不但能夠打破傳統馮紐曼架構下的記憶體瓶頸，同時能夠實現乘法與加法的平行化運算，藉此大幅度提升整體運算效能。

然而，由於記憶體內運算之記憶體需要額外的資料轉換介面，其包含數位類比及類比數位轉換器等，這些類比元件的效能會大大影響整體電路的吞吐量、能量消耗與面積使用效率，使得記憶體內運算之記憶體的效能受限，進而限制此種記憶體架構的應用時機。

本作品提出了一個應用於終端AI設備之高吞吐量、高能量與面積使用效率之記憶體內運算之靜態隨機存取記憶體（CIM SRAM）。本作品藉由改善資料處理與轉換電路，藉此克服CIM SRAM目前在性能上所受到的限制。

本作品提出了（1）動態加權二進制之數位類比轉換器，利用此電路架構改善先前電路需要額外能量消耗及運算線性度受限的問題，藉此提高整體記憶體之運算速度、能量使用效率與線性度。同時利用（2）所提出之統一式電荷運算網路，能夠改善電荷運算之訊號處理路徑，藉此提高類比訊號處理及資料轉換的能量與面積使用效率。本作品量測之晶片是實作於28nm製程，藉由所提出之設計，此CIM SRAM設計能夠達到186.16 GOPS的高吞吐量、41.87 TOPS/W的高能量使用效率，以及3288.4 GOPS/mm^2的高面積使用效率，相較於過去之文獻，在吞吐量、能量使用效率、面積使用效率上分別都有2.26倍、1.12倍及2.89倍的提升。



▲ 圖一 馮紐曼架構與記憶體內運算架構之比較

指導
教授

劉宗德 臺灣大學電機工程學系

美國加州大學柏克萊校區電機與計算機科學博士，現為臺灣大學電子工程學研究所 / 電機工程學系副教授。擔任教職之前，曾於聯發科技及比利時校際微電子研究中心從事積體電路設計與研發工作。

研究領域

應用於智慧終端裝置之低功率電路設計與系統開發

指導
教授

闕志達 臺灣大學電機工程學系

美國加州理工學院電機博士，現任臺灣大學電機工程學系教授。曾任國家晶片系統設計中心主任，國家實驗研究院副院長。

研究領域

基頻通訊積體電路設計、神經網路積體電路設計、無線通訊系統

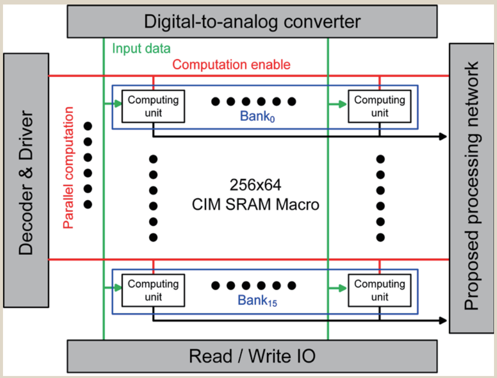
Abstract

With the advancement of artificial intelligence (AI) and machine-learning (ML) technologies, neural network-based ML architectures have demonstrated outstanding classification performances for many applications such as image and speech recognition. Accompanied by the rising of IoT products, the ML structures are massively applied on local edge devices for better performance. Compared to traditional cloud AI computing, edge AI computing can realize lower latency with better efficiency. Moreover, it achieves higher reliability and privacy by circumventing the need for the Internet connection. However, edge devices have substantially limited energy and computation resources. This makes the implementation of complex ML structures on a resource-constrained edge device really challenging.

For the emerging edge AI applications, computing-in-memory (CIM) architecture is proposed to further enhance the computation efficiency when performing ML tasks. The CIM architecture obviates the expensive data movement by performing computation directly inside the memory. This architecture achieves substantially higher energy efficiency than the conventional Von Neumann computation architecture, whose energy consumption is dominated by data movement.

However, since the implementation of a CIM SRAM requires additional data conversion interfaces, including digital-to-analog converters (DACs) and analog-to-digital converters (ADCs), the performances of these analog circuits ultimately limit the throughput, energy, and area efficiency of a CIM SRAM. This problem seriously constrains the application space of CIM SRAM to AI applications with low or moderate performance requirements.

In this work, we present a high-throughput, energy-area-efficient CIM SRAM for resource-constrained edge AI applications. This work overcomes the traditional performance bottleneck of CIM SRAM resulting from the data processing and conversion circuits by employing (1) the proposed dynamic current-steering DAC that achieves high conversion speed, energy efficiency, and linearity performance, and (2) the proposed unified charge processing network that simultaneously provides analog signal processing and data conversion functions with high energy and area efficiency. A test chip was fabricated in 28-nm CMOS technology to verify the proposed CIM SRAM design, which achieves a high throughput of 186.18 GOPS, with energy and area efficiency of 41.87 TOPS/W and 3288.4 GOPS/mm^2. This represents 2.26×, 1.12×, and 2.89× performance improvements in throughput, energy, and area efficiency, respectively, compared to the state-of-the-art high-performance CIM SRAM designs.



▲ Fig. 2 System architecture of the proposed CIM SRAM