D25-110

彈性且高效之記憶體內運算處理器

Flexible and Efficient CIM-based Processor

隊伍名稱｜杯子蛋糕是馬芬 Cupcake Is Muffin

隊　　長｜戴瑋佑 / 臺灣大學電子工程學研究所

隊　　員｜曾維雋 / 臺灣大學電子工程學研究所

魏振宇 / 臺灣大學電子工程學研究所

指導教授

劉宗德｜臺灣大學電機工程學系暨電子工程學研究所

臺灣大學電機工程學士、碩士、美國加州大學柏克萊校區電機與計算機科學博士，現為臺灣大學電機工程學系、電子工程學研究所教授。擔任教職之前，曾於聯發科技參與無線通訊電路與系統設計、於比利時校際微電子研究中心從事積體電路技術研究。

研究領域

高能效電路與系統設計。

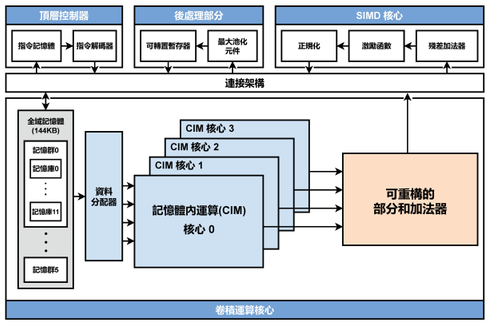
作品摘要

記憶體內運算（Computing-in-memory）技術已被廣泛研究，以實現具能源效率的神經網路（Neural Network）處理器。然而，多數數位CIM採用位元序列（bit-serial）的輸入方式，需要多個時鐘週期來累加部分和（partial sum），導致延遲時間與輸入位元寬成正比，進而降低面積效率。此外，加法器的實作通常是數位CIM巨集中主要的面積與功耗來源。有些文獻引入額外的加法電路來提升吞吐量，但由於增加了額外的加法與處理有號數運算的功能，這些技術反而加劇了加法器的問題，僅帶來有限的面積效率改善。

其次，儘管CIM巨集本身能達到高效能，但在整體CIM處理器的效能提升卻受到時間與空間使用率低落的限制。雖然有研究提出乒乓CIM巨集（ping-pong CIM macro），透過同時進行計算與記憶體更新操作來提升時間使用率，但其效益仍受限於CIM效能與記憶體寫入頻寬之間的平衡。此外，不同神經網路模型層與層之間的配置差異也嚴重限制了處理器的空間使用率。

最後，在傳統的CIM處理流程中，仍存在大量晶片內的權重與輸入/輸出資料轉移，這大幅增加了整體能耗並嚴重降低系統層級的能源效率。

為克服這些挑戰，本作品提出了：（1）平行輸入CIM巨集架構，與傳統位元序列輸入CIM相比，實現了4倍的吞吐量與2.4倍的面積效率提升。（2）彈性資料流與核結構，與傳統CIM架構相比，執行週期與能耗分別降低了64.8%與37.7%。（3）多功能資料緩衝器，進一步降低記憶體存取59.1%與能耗26.2%。本作品將所提出的記憶體內運算處理器實作在28奈米CMOS製程，晶片面積為2.3mm^2。在60MHz和0.6V的操作電壓下，晶片的消耗功率為21.6mW。與最先進的記憶體內運算處理器相比，所提出的處理器在系統面積使用效率和效能指標上分別提升了3.87倍和1.36倍，展示了其在面積使用效率和能量使用效率上的綜合競爭力。



圖一 系統架構圖。

Abstract

Computing-in-memory (CIM) has been extensively explored to realize energy-efficient neural network (NN) processors. However, the bit-serial input scheme utilized by most digital CIMs requires multi-cycle partial sum (psum) accumulation, causing prolonged latency proportional to input bit-width and degraded area efficiency. Moreover, the adder implementation overhead often dominates the area and power of digital CIM macros. Some works introduce an additional adder circuit to enhance throughput, but these techniques exacerbate the adder cost issue due to the additional addition and sign-extension functions, resulting in only marginal area efficiency improvement.

Secondly, although CIM macro can achieve high performance, the corresponding performance improvement in the CIM processors diminishes due to low temporal and spatial utilization. While ping-pong CIM macro is proposed to enhance temporal utilization with simultaneous compute and memory updating operations, its effectiveness is largely constrained by the balance between workload, CIM performance, and memory write bandwidth. On the other hand, distinct configuration across different NN models and layers severely limits the processor spatial utilization. Finally, a significant amount of on-chip weight and input/output data movement still exists in the conventional operational flow of CIM processors, substantially increasing the overall energy consumption and greatly degrades system-level energy efficiency.

To overcome these challenges, this work proposes: (1) A bit-parallel input CIM macro architecture, achieving 4× throughput and 2.4× area efficiency improvement compared to conventional bit-serial input CIM. (2) A flexible dataflow and core topology, reducing the overall execution cycles and energy consumption by 64.8% and 37.7%. (3) A versatile data buffer further realizing a 59.1% and 26.2% reduction in memory access and energy.

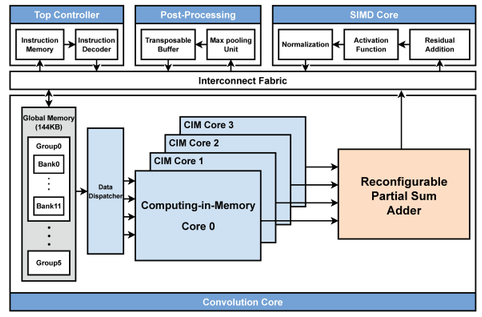


Fig. 2 System structure.