Design Group

D17-030

應用於非揮發性處理器之記憶體內運算電路與非揮發性邏輯電路

Energy-efficient Process in Memory Circuit and Nonvolatile Logic Used in Nonvolatile Processor

隊伍名稱 攻核機動隊
Ghost in Core
隊　　長 陳韋豪 / 清華大學電子工程研究所
隊　　員 徐國翔 / 清華大學電子工程研究所
李峻毅 / 清華大學積體電路設計與製程開發產業碩士專班
邱曉昀 / 清華大學電機工程學系

指導教授
張孟凡
清華大學電機工程學系

成功大學電機工程學士、美國賓州州立大學電機工程碩士、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、記憶體內運算電路設計、人工智慧晶片之記憶體電路設計。

作品摘要

隨著 IoE 裝置與無線傳輸系統發展，低功耗處理器成為熱門技術開發。然而，IoE 裝置應用系統低電容量與使用環境能源不穩定問題，其系統需具長待機時間與低耗能的特性。新興式記憶體（Emerging memory），如電阻式記憶體（ReRAM）、相變化記憶體（PCM）與磁性記憶體（MRAM）...等具有非揮發性、高密度、低功耗等特性。適合應用於非揮發處理器當中，用以備份處理器待機時資訊，並在復機後將資訊重新使用，大幅減少傳統揮發性（Volatile）記憶體待機時的漏電問題。

傳統處理器為馮諾伊曼（Von Neumann）架構，運算資料於處理器與記憶體巨集間傳遞，此架構速度受限於資料傳輸介面的輸入輸出端（IO）數，且搬動資料需要消耗大量額外能量，稱為「Memory-wall」。近年，類神經網絡架構（Neural network）成為熱門的研究項目。有別於傳統馮諾伊曼架構，類神經網絡可於單晶片中實現平行運算，達到快速且低功耗之運算目標。

利用高密度、高低阻態比值（R-ratio）大的電阻式記憶體（ReRAM），本研究成功設計應用於非揮發性處理器之類神經網絡之記憶體內運算（Processing-In-Memory）電路與非揮發性邏輯（nvLogic）電路。下列為本研究之電路特色：

1. 國際首次，具高度整合性之記憶體內運算電路與非揮發性邏輯電路並應用於實際微處理器。
2. 記憶體內運算電路開發，作為處理器之加速單元，可有效降低計算功耗與速度提升。
   - 國際首次，高度整合之記憶體內運算電路晶片實作。
   - 利用類神經網絡架構可於記憶體巨集內實現平行運算，無須將資料傳遞至處理器再進行運算。
3. 非揮發性邏輯電路開發，可於單巨集中對資料進行平行備份與讀取，達到快速、低耗能開關機特色。
   - 創新的自動寫入截斷機制（Self-Write-Termination）可大幅降低資料備份耗能。
   - nvFF 可用以取代處理器內關鍵運算單元暫存器，使得處理器可在能源不穩定時持續進行開機步驟，不須重新執行。

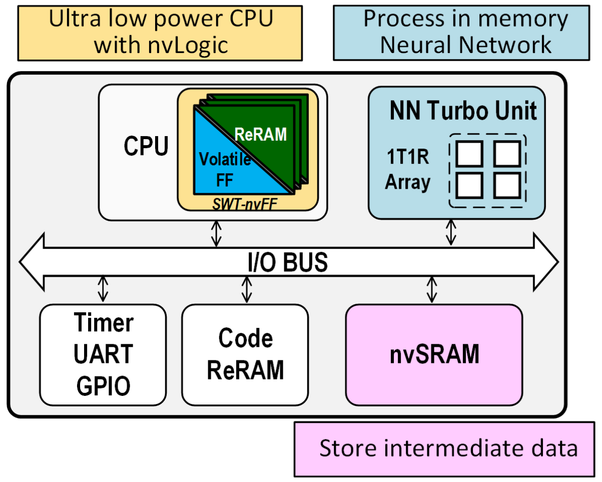

圖 1. 結合記憶體內運算與非揮發邏輯電路之高效能非揮發性處理器架構圖

Abstract

With the ever-increasing market demands of internet of everything （IoE） and wireless sensor networks （WSN）, low energy processors have become a great topic of interest. In particular, the end-devices must support long idle time and low energy consumption due to the limited and noisy power supply.

Emerging memories, such as ReRAM, PCM and MRAM are suitable for nonvolatile processor （nvProcessor） due to its nonvolatility, high density, and low power consumption. It can be used to backup data before power-off and recover data once the power is restored. This operation greatly reduces the leakage energy of volatile memories during standby.

Conventional processors employ the Von Neumann architecture, where data processing requires transmission between CPU and memory. The throughput of this approach is limited by the IO interface, and transmitting data consumes a large amount of excess energy. This is known as the “Memory-wall”. Recently, neural networks which differ from the conventional architecture has become a popular topic due to its ability to conduct efficient parallel computations, enabling high speed and low power data processing.

By using 150nm CMOS process and HfOx RRAM, this research successfully demonstrates an energy-efficiency nonvolatile processor with a novel Self-Write-Termination 2R-nvFF （SWT-2R-nvFF） and RRAM based Processing-in-memory （PIM） circuits. The main contribution of this work is as the follows:

1. For the first time, RRAM based PIM circuits and nonvolatile logics （nvLogic） have been integrated in a nonvolatile processor.
2. Processing-in-memory circuits serve as the accelerator for a high efficiency processor.
   - For the first time, a RRAM based processing-in-memory integrated chip is presented.
   - The PIM circuitry supports neural network computations through parallel processing in memory, eliminating data transmission to the CPU.
3. Nonvolatile logics allow parallel, local backup and recovery of data to reduce energy consumption and enable fast power-on-off operations.
   - A new Self-Write-Termination nvFF （SWT-nvFF） greatly reduces data backup energy.
   - NvLogic make it possible for the processors to continue computation even with rapid power interrupts by storing critical data to NVM.

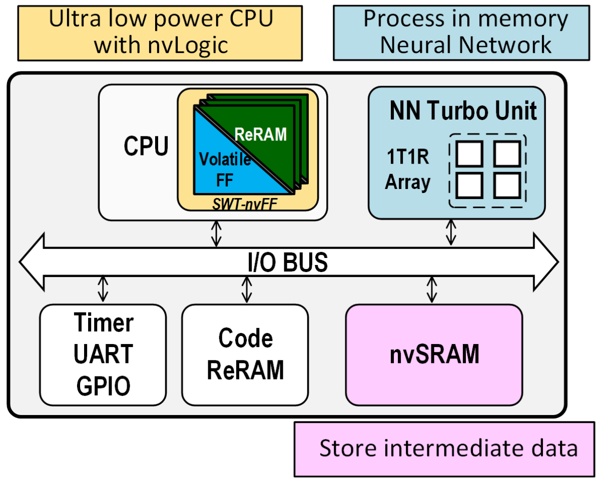

Fig 2. High energy efficiency nonvolatile process with processing-in-memory circuit and nvLogic