A26-183

具後段製程相容性之3D異質整合2T eDRAM：應用於邏輯內記憶體之高效能快取架構

3D Heterogeneous Integrated
BEOL-Compatible 2T eDRAM Cache
for Memory-in-Logic Applications



隊伍名稱｜快取混搭隊 Hybrid Cache Team

隊　　長｜林靖萍 / 清華大學電機工程學系

隊　　員｜顏維萱 / 清華大學電機工程學系

隊　　員｜張正昕 / 清華大學電機工程學系

隊　　員｜鐘健誠 / 清華大學電機工程學系

指導教授 ADVISOR



葉昭輝 | 清華大學電機工程學系

服務於電機工程學系、電子工程研究所，並合聘於清華大學半導體研究學院。積極推動學術研究與產業應用接軌，致力培育具備半導體元件知識技術與機器學習硬體晶片設計跨領域能力之人才。

研究領域

聚焦於二維半導體電子元件(2D Electronics)、積層型3D異質積體電路整合技術(Heterogeneous M3D-Ics)、嵌入式記憶體(Embedded Memory)、In-Memory Computing、超高維度運算(Hyperdimensional Computing)，以及 AI 硬體加速器(AI Accelerators)等前瞻技術研究。

作品摘要

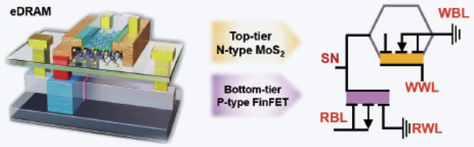
隨著人工智慧(AI)、大型語言模型(LLM)及高效能運算(HPC)等資料密集型應用快速發展，處理器與記憶體之間頻繁的資料傳輸已成為限制系統效能與能源效率的重要瓶頸，形成所謂的「記憶體牆(Memory Wall)」問題。傳統快取記憶體主要採用SRAM架構，雖具有低延遲特性，但其單元面積大且靜態功耗高；另一方面，DRAM具備高儲存密度，卻因需依賴實體電容與頻繁刷新機制而增加功耗與製程複雜度。因此，如何發展兼具高密度、低功耗及高整合度的新型嵌入式記憶體技術，已成為未來晶片設計的重要研究方向。

本研究提出一種具後段製程相容性(BEOL-compatible)之三維異質整合2T eDRAM 快取架構，應用於Memory-in-Logic系統。該架構利用無電容式兩電晶體(2T)增益單元取代傳統1T1C eDRAM結構，透過電晶體閘極與寄生電容儲存資料，避免實體電容在先進製程微縮下所面臨的限制。研究中採用頂層二維材料MoS_2 nFET與底層矽基p-type FinFET進行單體三維異質整合(Monolithic 3D Integration)，並結合HfO_2高介電常數閘極絕緣層及低於400°C的低熱預算製程技術，使記憶體單元可直接建構於邏輯電路上方，達成高密度垂直整合之目標。

實驗結果顯示，所開發之MoS_2元件具有極低關態漏電流(低於10^{-14}A)及超過10^8的開關比，可有效提升資料保持能力。所實現之2T eDRAM具備約10 ns的高速寫入能力，資料保持時間可達10^2至10^3秒以上，相較傳統矽基eDRAM提升約六個數量級，大幅降低系統刷新頻率與待機功耗。Memory-on-Logic架構及下世代高效能運算系統提供了一項具前瞻性與實用價值的解決方案。

此外，本研究成功驗證多階儲存操作能力，可實現多個穩定且可區分之儲存狀態，展現高資訊密度儲存潛力。陣列量測結果亦顯示良好的單元一致性與可靠度，證實其具備大規模整合應用之可行性。

在系統層面，本研究利用M3D技術將記憶體直接堆疊於邏輯層上方，可有效縮短資料傳輸距離，降低延遲與能量消耗。相較於傳統平面式架構，整體邏輯面積可縮減約40%，顯著提升晶片整合密度。研究成果不僅可應用於高效能運算與人工智慧晶片之嵌入式快取記憶體，亦具備延伸至邊緣運算、低功耗物聯網及存算一體(Computing-in-Memory)系統之潛力。透過結合新材料、低溫製程與三維異質積體電路整合技術，本研究為未來Memory-on-Logic架構及下世代高效能運算系統提供了一項具前瞻性與實用價值的解決方案。



圖一 M3D 異質整合 2T eDRAM 元件結構及操作電路示意圖。



圖二 eDRAM array 與單顆 cell 之光學影像圖。

Abstract

With the rapid advancement of artificial intelligence (AI), large language models (LLMs), and high-performance computing (HPC), the demand for memory capacity and data bandwidth continues to increase. Frequent data transfer between processors and memory has become a major bottleneck in modern computing systems, commonly known as the "memory wall." While SRAM provides low-latency access, its large cell area and high static power consumption limit scalability. DRAM offers higher density but suffers from refresh overhead and integration challenges. Therefore, developing a high-density, low-power, and highly integrated embedded memory architecture has become a critical challenge for next-generation computing systems.

This work proposes a BEOL-compatible three-dimensional heterogeneous integrated 2T eDRAM cache architecture for Memory-in-Logic applications. The proposed design adopts a capacitorless two-transistor (2T) gain cell instead of the conventional 1T1C eDRAM structure. Data are stored through transistor gate capacitance and parasitic capacitance, eliminating the need for physical capacitors. The architecture vertically integrates a top-tier MoS_2 nFET with a bottom-tier silicon p-type FinFET using monolithic 3D integration technology. Together with an HfO_2 high-k dielectric and a low-temperature process below 400°C, the memory layer can be directly integrated above logic circuits while maintaining BEOL compatibility.

The fabricated MoS_2 transistors demonstrate excellent electrical characteristics, including an ultra-low off-state leakage current below 10^{-14} A and an on/off current ratio exceeding 10^8. These properties significantly enhance charge retention capability and reduce standby power consumption. The proposed 2T eDRAM achieves a write latency of approximately 10 ns and a retention time ranging from 10^2 to 10^3 seconds, representing an improvement of nearly six orders of magnitude compared with conventional silicon-based eDRAM. The extended retention time substantially lowers refresh frequency requirements and reduces refresh-related energy overhead. In addition, the memory cell supports multiple distinguishable storage states, indicating strong potential for multi-level storage. Array-level measurements further confirm good uniformity and reliability, demonstrating the feasibility of large-scale integration for future memory-on-logic platforms.

From a system perspective, the proposed M3D architecture enables memory cells to be stacked directly above logic circuits, significantly reducing interconnect distance, latency, and data movement energy. Compared with conventional planar implementations, the architecture achieves approximately 40% area reduction while maintaining high performance. The proposed technology shows strong potential for embedded cache memory in AI accelerators, HPC processors, and advanced system-on-chip platforms. Furthermore, its low-power and high-density characteristics make it suitable for edge computing, IoT devices, and computing-in-memory applications. This work provides a promising pathway toward future Memory-on-Logic systems and next-generation high-performance computing architectures.

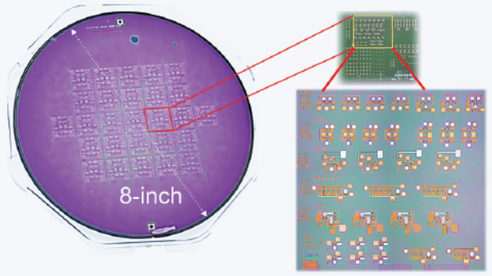


Fig. 3 Chip and structure images of the M3D heterogeneous integrated devices fabricated on an 8-inch wafer.

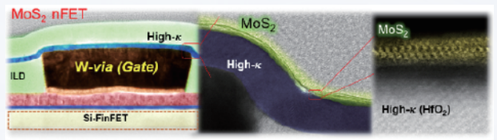


Fig. 4 Cross-sectional TEM image of the vertically stacked MoS_2 nFET, high-k HfO_2 dielectric, and Si FinFET structure.