Defect-tolerant tree-based models in MoS_2-based Flash memory arrays with Sb contact

Bo Wen, Guoyun Gao, Can Li
Department of Electrical and Electronic Engineering,
The University of Hong Kong, Pokfulam, Hong Kong SAR
canl@hku.hk

The rapid advancement of artificial intelligence has raised concerns about trustworthiness, particularly regarding interpretability and robustness. While tree-based models like Random Forest and XGBoost provide superior interpretability and accuracy for tabular data, their scaling remains computationally expensive due to poor data locality and high data dependence. Previous attempts to accelerate these models using analog content addressable memory (CAM)[1] have been unsuccessful because sharp decision boundaries are both difficult to implement and highly vulnerable to device variations, resulting in compromised hardware performance and susceptibility to adversarial attacks[2].

This work introduces a novel hardware-software co-design approach that leverages MoS_2 Flash-based analog Content addressable memories (CAMs) with inherent soft boundaries, enabling efficient inference with soft tree-based models[3]. While CAMs offer an innovative solution by processing information directly within memory, conventional implementations using SRAM and emerging nonvolatile memories remain limited by silicon transistor performance. To address this limitation, we developed an ultra-compact analog CAM utilizing atomic thin 2D MoS_2 flash memories with semimetal antimony (Sb) contacts. Our device demonstrates unprecedented performance with the highest readout current (60 μA/μm) and large ON/OFF ratios (>10^9) reported in 2D flash memories, achieving record low energy consumption and latency (below 0.1fJ/10ps per search per cell) during analog in-memory search operations in our physical 8x16 analog CAM array with 256 MoS_2 flash memory devices.

Experimental results confirm that our soft tree model inference on MoS_2 analog CAM arrays delivers exceptional robustness against both device variation and adversarial attacks while maintaining state-of-the-art accuracy. Our fabricated analog CAM arrays achieve 97% accuracy on the Iris dataset, and our experimentally calibrated model demonstrates remarkable resilience, showing only a 0.6% accuracy drop on the MNIST dataset under 10% device threshold variation, compared to a substantial 45.3% drop for traditional decision trees. This breakthrough paves the way for specialized hardware that significantly enhances both the trustworthiness and efficiency of AI systems.

---

[1] C. Li, *et al.*, Nat Commun 11, 1638 (2020).
[2] G. Pedretti, *et al.*, Nat Commun 12, 5806 (2021).
[3] O. Irsoy, et al, ICPR 2012

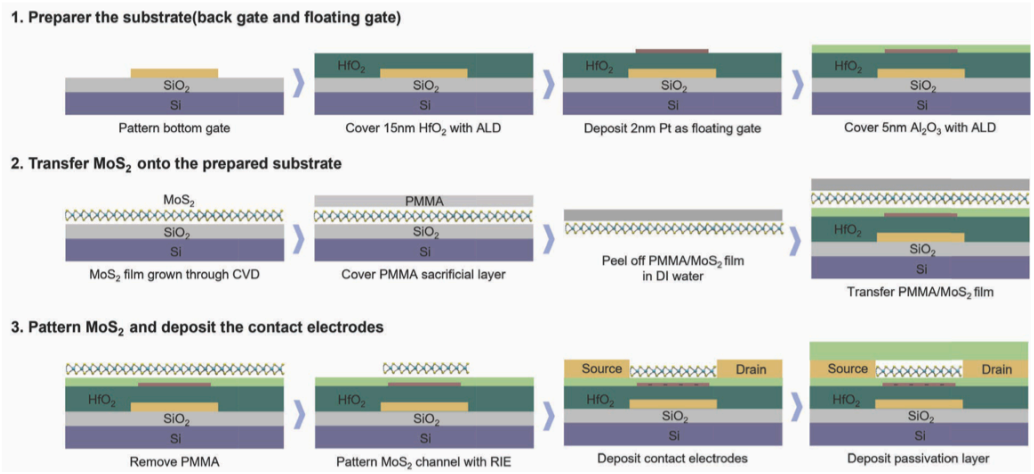

Figure 1: Fabrication process of our MoS2-based Flash memory devices

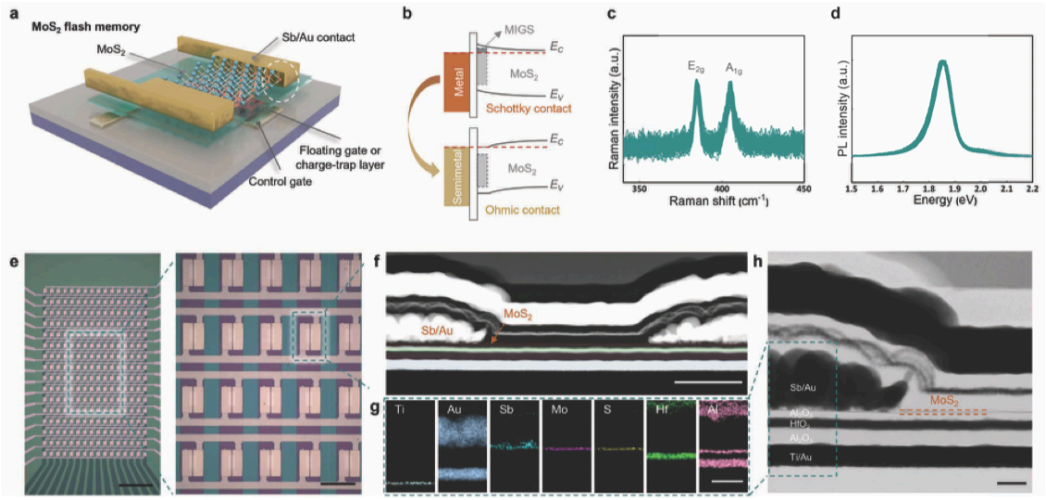

Figure 2: Our MoS2-based Flash memory analog CAM arrays

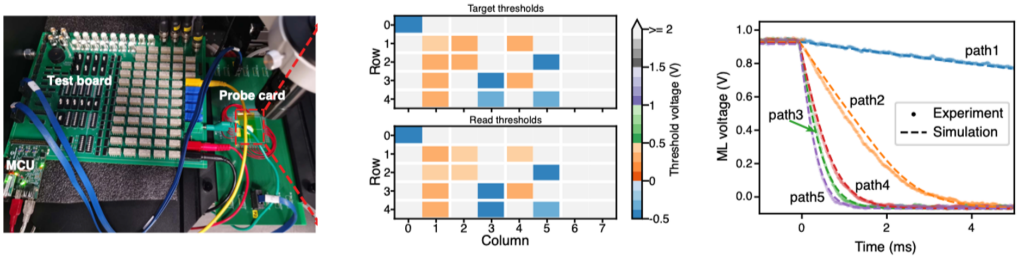

Figure 3: Experimental results of soft tree model inference with our custom-built measurement system and MoS2-based Flash memory array