Using Non-Volatile Memory for Computation-in-Memory

Wei -Ti Liu

LucidPort Technology, Inc.

www.lucidport.com

Using Non-Volatile Memory for Computation-in-Memory

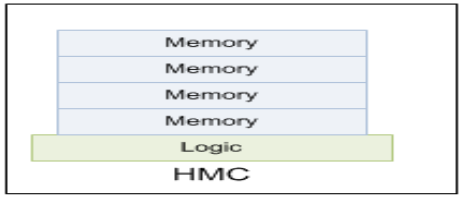
- Non-volatile memory offers major opportunities for computation-in-memory (CIM). With the recent increasing interest in artificial intelligence (AI), academic researchers and startup companies are exploring the use of non-volatile memory technology (floating gate, MRAM, RRAM, and 3D Xpoint) to support CIM for AI applications. We explore the related design and manufacturing issues for various non-volatile technologies.
- When CIM performs machine learning, the non-volatile memory is subject to frequent write cycles. The analog circuits used for machine learning must be directly coupled with an array of non-volatile memory cells. Analog circuits impose much more stringent requirements on signal accuracy and integrity, resulting in much less tolerance for small defects or bugs. Thus, even minor design modifications can create design and implementation problems which will affect the testing, reliability, manufacturing, and final cost of products.

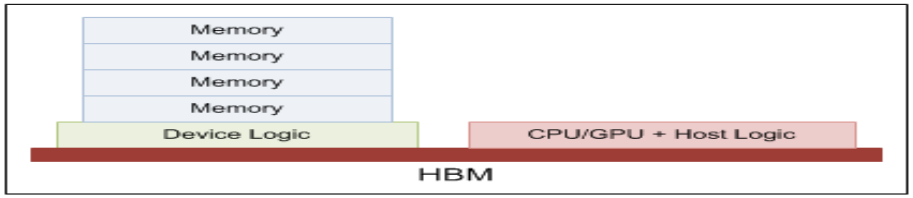
Background
Computation-in-Memory

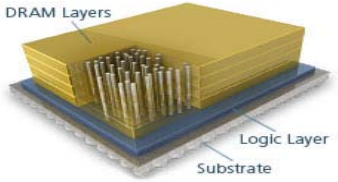
- Processing-in-Memory can take advantage of the logic unit merged with 3-D stacked memories
- Processing-in-Memory avoids the frequent need to move data from memory to CPU.
- Moving a portion of computation to memory takes advantage of memory's high internal bandwidth

Processing-in-Memory (PIM)

- PIM Today: High Cost; Used in High End Systems
- HMC/HBM based CIM
- TSV (Through Silicon Vias)







Source: Micron

- High bandwidth
- Low power consumption
- High density
- Moving part of computation to where data resides (memory)

Ref: Erfan Azarksh, Davide Rossi, Igor Loi, and Luca Benini
Design and Evaluation of a Processing-in-Memory Architecture for the Smart Memory Cube

Ref: Hybrid Memory Cube Specification 1.0
Hybrid Memory Cube Consortium

Flash Memory Summit

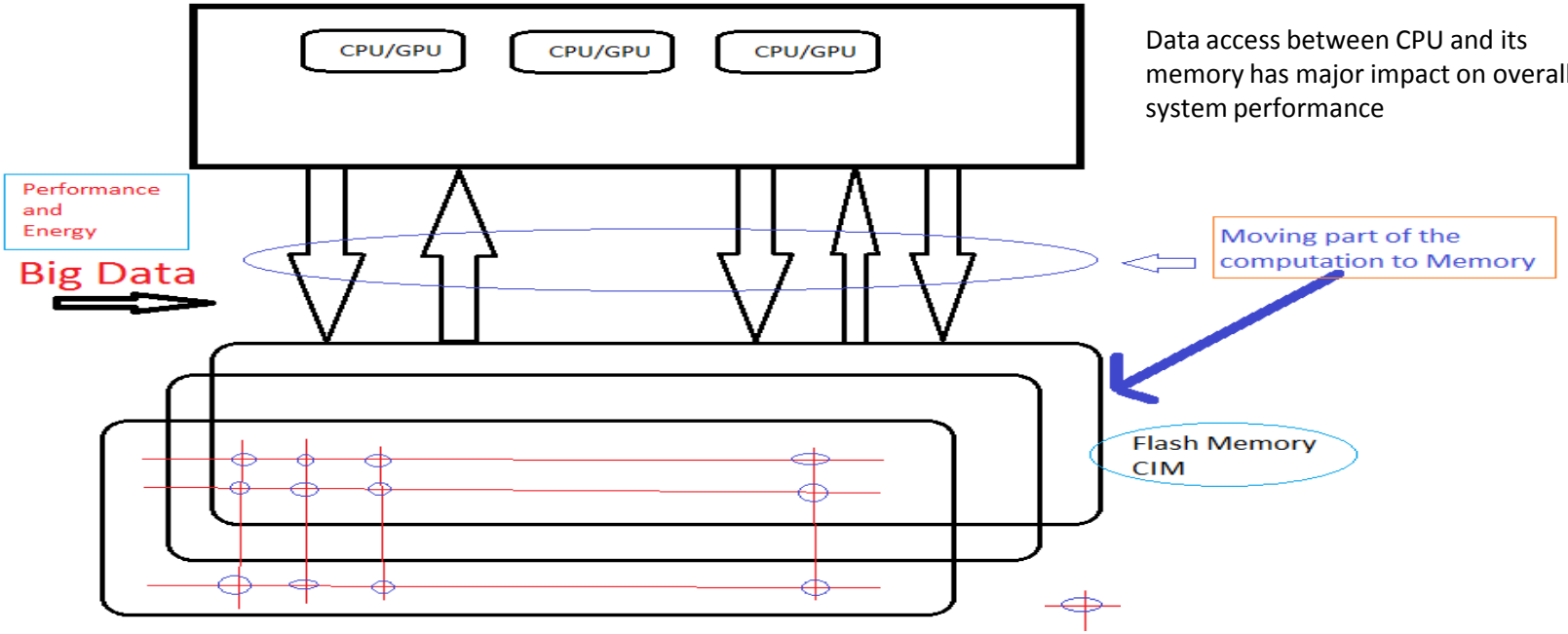
Computation-in-Memory(CIM)

- Companies are exploring Non-volatile memory technologies to implement Computation-in-Memory for AI applications
  - Perform Analog Computer functions
  - Computation is in Flash memory—Fast
- Advantages of NVMs
  - Non-volatile property, high density
  - Low power and lower cost

Ref: Jintao Zhang, Zhuo Wang and Naveen Verma

In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array,

IoTs Big Data Processing

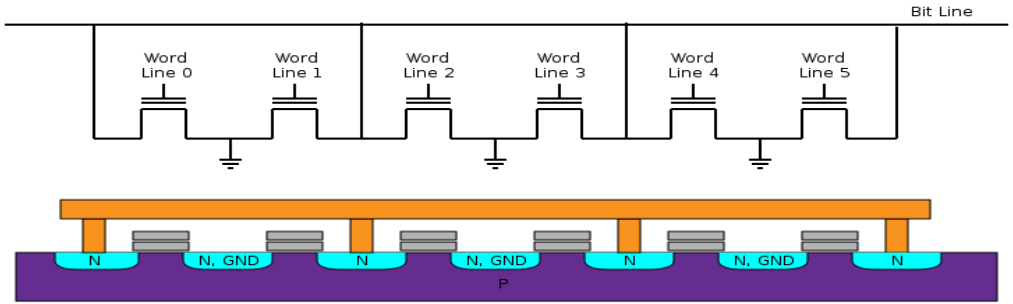


Data access between CPU and its memory has major impact on overall system performance

Flash memory Cell

Using NOR Flash Memory for CIM

NOR-- basic operation



Source: https://en.wikipedia.org/wiki/Flash_memory

Computation-in-Memory(CIM) (addition/multiplication)

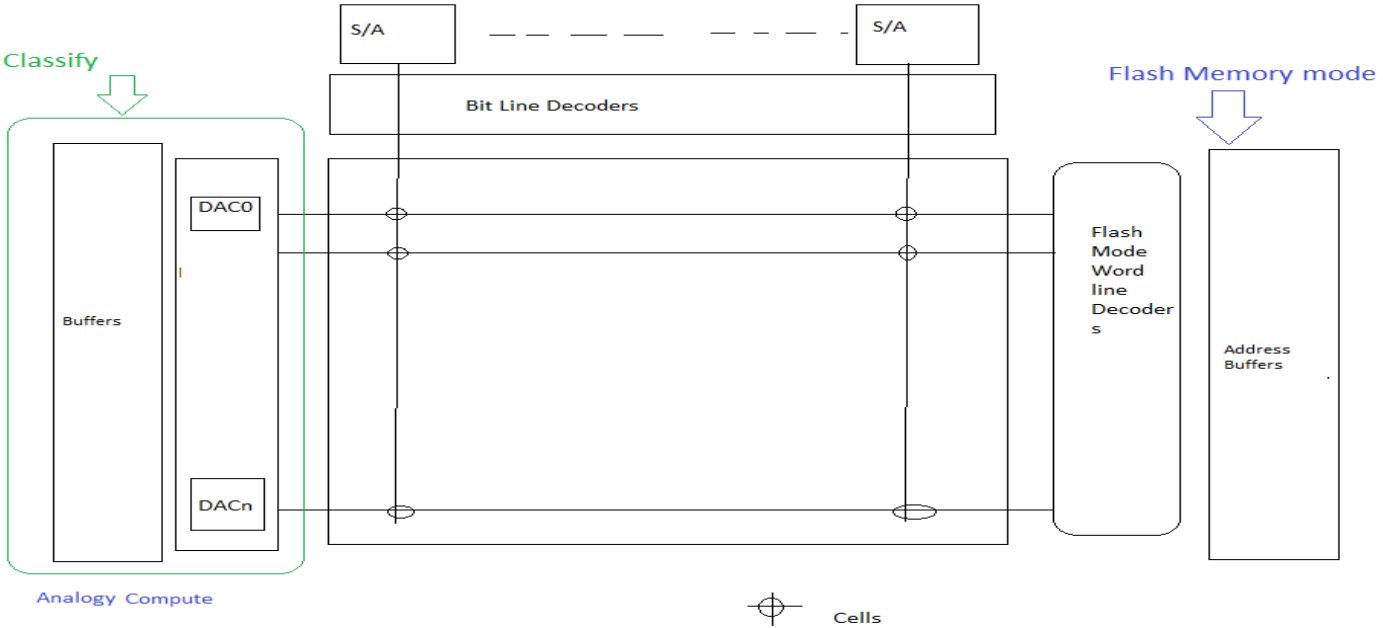
- Support in-Memory multiplication and enables in-memory addition/multiplication
  - Using multiple NOR cells and NOT gate
- NOR Operation is the basic cell
  - Multiplication/addition
  - OR, AND and XOR etc.
- No need to make changes on NVM major circuits
  - Sense Amplifiers

Using Flash Memory for CIM Circuit Design

- Flash memory Mode
  - Flash memory design performs typical Flash memory functions
  - Read/Write
  - Storing data
- Classify Mode

Flash Memory Summit

Block diagrams Using NVMs for CIM



Non-volatile memories for CIM

- Floating Gate Flash memory
- Emerging NVMs
  - MRAM
  - 3D Xpoint
  - Others
    - ReRAM, FRAM

Ref: Xiangyu Dong, Xiaoxia Wu, Guangyu Sun, Yuan Xie , Helen Li and Yiran Chen ‘ Circuit and Micro architecture Evaluation of 3D Stacking Magnetic RAM(MRAM) as a Universal Memory Replacement
Jason Heidecker, Jet Propulsion Laboratory, Pasadena, California

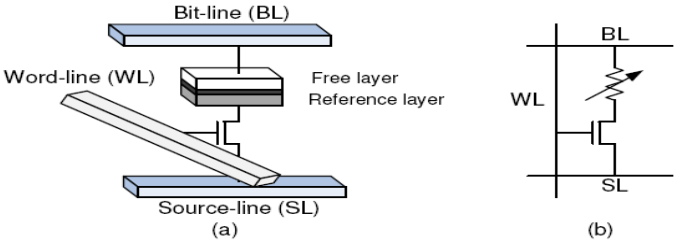
CIM Flash Memory (Floating Gate)

- Flash memory limitations
  - Wear-out
    - Raw Bit error rate (RBER)
    - RBER grows exponentially as P/E cycles increase
  - Block erasure
  - Slow write
- P/E (Program / Erase) cycles
- Random access
- Mature technology, CMOS compatible
  - Cost

Flash Memory Summit

MRAM

- STT-MRAM (Spin-transfer Torque MRAM):
  - Spin* is transferred from conduction electrons to a non-volatile magnetized layer
- Embedded Flash
  - Good data retention
  - Good endurance
- Fast speed-Read
- CMOS compatible
- Manufacturing?
  - Higher cost and manufacturing issues
- Stray Magnetic Fields



*Spin is one of two types of angular momentum in quantum mechanics first proposed for electrons in the 1920s. Electron spin can often be accurately visualized as a very small rotating current. Source: https://en.wikipedia.org/wiki/Spin_(physics)

3D Xpoint

- 3D Xpoint is Phase-Change-Memory
- Individual cells do not need a transistor and high density
- Fast Read/Write
- Good endurance, million write cycles
- 3D Xpoint's cost is in between DRAM and NAND Flash
- Not well architected for In-memory-processing?
- Limited suppliers

Summary

- By using NVMs for AI application, we have to resolve many challenges faced by designers in circuit design and circuit/ architecture interactions
- To adopt NVMs for CIM applications, we have to improve the trade off between performance and power consumption
- The NVMs require on-chip analog circuits to perform deep learning and artificial intelligence (AI) applications. This results in increased manufacturing cost and energy
- Emerging NVMs present higher manufacturing cost than conventional memory.
- However, if these issues can be addressed or mitigated, system level performance could potentially be increased 10X.

Reference

1. Jintao Zhang, Zhuo Wang and Naveen Verma
   In-Memory Computation of a Machine-Learning Classifier in a Starndard 6T SRAM Array,
2. Sophiane Senni, Lionel Torres, Gilles Sassatelli, and Abdoulaye Gamatie, LIRMM, UMR CNRS 5506, University of Montpellier
   Non-Volatile Processor Based on MRAM for Ultra-Low-Power IoT Devices
3. T.M. Maffitt et al.IBM Res & DeV. Vol. 50 NO 1 January 2006
   Design Considerations for MRAM
4. MRAM Technology Status
   Jason Heidecker, Jet Propulsion Laboratory, Pasadena, California JPL Publication 13-3 2/13
5. Mohsen Imani, Saransh Gupta, Tajana S. Rosing, University of California San Diego, System Energy Efficiency Lab
   Accelerating Multiplication and Parallelizing Operations in Non-Volatile Memory
6. Erfan Azarkhish, Davide Rossi, Igor Loi, and Luca Benini, Springer International Publishing Switzerland 2016
   Design and Evaluation of a Processing-in-Memory Architecture for the Smart Memory Cube
7. SAUGATA GHOSE, KEVIN HSIEH, AMIRALI BOROUMAND,RACHATA AUSAVARUNGNIRUN Carnegie Mellon University, ONUR MUTLU
   ETH Zürich and Carnegie Mellon University
8. Enabling the Adoption of Processing-in-Memory: Challenges, Mechanisms, Future Research Directions
   Jin Hee Cho, et al. SK hynix, Gyeonggi, Korea, 2018 IEEE International Solid-State Circuits Conference
9. A 1.2V 64Gb 341GB/s HBM2 Stacked DRAM with Spiral Point-to-Point TSV Structure and Improved Bank Group Data Contro
10. 3D-Xpoint, Intel
    https://www.intel.com/content/www/us/en/architecture-and-technology/optane-technology-animation.html