Amirkabir University of Technology
(Tehran Polytechnic)

Memory Technologies Course By

Dr. Hammed Farbeh

Homework 3

CE5431 | Spring 2024

Teaching Assistants

Morteza Adelkhani (Madelkhani@aut.ac.ir)

Sarah Zamani (sara.zamani73@aut.ac.ir)

Description:

Currently, one of the most popular concepts in computer architecture is processing in memory. Hence, in this homework, we want to delve into this area and conduct some implementation and research.

Processing in memory (PIM) is a computing architecture that brings computation closer to where data is stored, typically in the memory subsystem. This approach is designed to minimize data movement between the CPU and memory, thereby reducing latency and power consumption associated with data transfer. PIM can be implemented in various ways, such as by adding processing capabilities directly to memory chips or by creating 3D stacked memory with integrated layers of processing elements.

PIM represents a shift from traditional computing paradigms by addressing the 'von Neumann bottleneck,' a limitation caused by separate storage and processing locations. As data volumes continue to grow, processing in memory is becoming an increasingly important innovation for efficient and fast data handling in a wide range of computing environments.

The integration of PIM for designing neural network accelerators is a transformative approach that addresses the critical challenges of data movement and computational efficiency in traditional architectures. Neural networks, particularly deep learning models, are inherently data-intensive and require substantial computational resources for tasks such as image recognition, natural language processing, and autonomous driving.

PIM architectures offer a paradigm shift by performing computations directly within the memory, thereby significantly accelerating the inference phase of neural networks. This is achieved through specialized memory architectures that support fast arithmetic operations like addition and multiplication, which are fundamental to neural network computations. By reducing the overhead of data movements between the memory and processing cores, PIM can lead to remarkable improvements in energy efficiency and speed, making it possible to achieve higher performance levels compared to traditional CPU or GPU-based systems.

Moreover, PIM enables weight sharing and parallel in-memory components, further enhancing computational speed and reducing energy consumption. This is particularly beneficial for edge computing devices and IoT applications where power efficiency is paramount. The ability to process large volumes of data quickly and efficiently within the memory also opens up new possibilities for real-time analytics and decision-making in various applications.

The use of non-volatile memory (NVM) in processing memory is gaining importance due to its ability to retain data even when the power is turned off. This characteristic is crucial for various computing applications, especially those requiring persistent data storage and quick access to data upon power-up.

In the context of PIM, incorporating NVM can enhance system performance and energy efficiency. NVM technologies like Phase-Change Memory (PCM), Resistive RAM (ReRAM), and Magnetoresistive RAM (MRAM) are being explored for PIM architectures. They offer the potential for faster access times and reduced power consumption compared to traditional volatile memory solutions such as DRAM.

Moreover, NVM can improve the endurance and reliability of PIM systems. Since NVM does not require power to maintain the stored information, it can significantly reduce the overall energy consumption of the system. This is particularly beneficial for battery-powered devices and edge computing applications where power availability is limited.

Additionally, NVM can enable new computing paradigms such as in-memory computing, where data can be processed directly within the memory array, leading to a reduction in data movement and latency. This is especially important for data-intensive tasks like neural network computations, where the speed of data access can greatly influence the overall processing time. **If you want to learn more about this concept you can read following papers.**

1- PRIME (ucsd.edu)
2- ISSAC (utah.edu)
3- TIMELY (arxiv.org)

Theoretical Questions:

Question 1:

A technology company is developing a neural network accelerator with a processing-in-memory (PIM) architecture for use in industrial automation systems. The PIM architecture allows computation to be performed directly within the memory devices, reducing data movement and improving energy efficiency. The company needs to decide between various memory devices, such as SRAM, DRAM, MRAM, and PCM, for implementing the PIM structure in the neural network accelerator.

In the context of our industrial neural network accelerator with a processing-in-memory (PIM) structure, compare the suitability of different memory devices (SRAM, DRAM, MRAM, PCM) for PIM implementation. Consider factors such as access speed, energy efficiency, endurance, area efficiency, and ease of integration. Provide a recommendation based on the specific requirements and constraints of the industrial automation application.

Based on the analysis, recommend the most suitable memory device (SRAM, DRAM, MRAM, PCM) for implementing the processing-in-memory (PIM) structure in the neural network accelerator for industrial automation. Justify your recommendation by aligning it with the specific requirements, performance goals, and constraints of the industrial automation application. Discuss potential trade-offs and any additional considerations that influenced your decision.

Implementation Questions:

* The technology file is attached in the homework folder.

* You have to attach your code and waveform figure in your response file.

Question 1:

By using HSPICE compare one STT-RAM cell as a candidate of nonvolatile memories with DRAM and SRAM cell. You can also use ReRAM instead of STT-RAM in this comparison, so this is optional. (**The technology file for STT-SHE-RAM is attached in the homework folder; you can use it for implementation in this homework.**)

* Note that you can choose one of the following questions and solve it.

Question 2:

Design a 3*3 bank memory based on a STT-RAM cell with a sense amplifier (SA) for reading data from the memory cell in Hspice and fill the corresponding rows in table 1 with STT-RAM. Finally, determine which memory is more suitable for designing main memory. Please append an output waveform to validate each STT-RAM cell. (as previous question you can use STT-SHE-RAM instead of STT-RAM)

Question 3:

set up NVSim on RAM mode, and must all memory types be evaluated for RAM design. After executing all tasks, analyze the differences in power, latency, and area. Finally, determine the best choice among all memory types for RAM, and thoroughly explain the reasons for this selection.

Table I

	Memory 1	Memory 2
Read time		
Write time		
Total Power		