Amirkabir University of Technology
(Tehran Polytechnic)

# Memory Technologies Course By

# Dr. Hammed Farbeh

**Homework 4**

**CE5431 | Spring 2024**

Deadline:2024.06.03

Teaching Assistants

Morteza Adelkhani (Madelkhani@aut.ac.ir)

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

**Description:**

In the previous class (PIM course), you were introduced to Processing-in-Memory (PIM) technology along with some examples in this field. In this homework, you will engage further with this concept.

**Theoretical Questions:**

Question 1:

a) What is PUM, and which types of memory are most commonly used for it? Please explain in detail why each type of memory is used.
b) What are the weaknesses of UPMEM?
c) Introduce the Ambit structure and its advantages and disadvantages.

**Implementation Questions:**

* You have to attach screenshot of your result for each output.

In this part of your assignment, you will be dealing with the MNSIM 2.0 simulator for implementing a neural network accelerator. Therefore, you must first download this simulator and provide the results as requested.

Basic Config:

1) Download the VGG8 parameters of the neural network that has been trained on the CIFAR-10 dataset, as described in the MNSIM 2.0 guideline PDF.

2) For each run, you must selectVGG8as your desired neural network.

Question 1:

As you learned in the class, the PIM structure consists of a number of tiles, and each tile includes a number of PEs. In each PE, we have some necessary circuits and a crossbar structure of memory cells. What happens if we reduce the size of the crossbar? For example, should we expect power, latency, and accuracy to decline, rise, or not change? Explain your answer in detail.

Question 2:

If we want to add PUM to this simulator, which part should be changed?

Question 3:

In the first implementation, set the size of the crossbar (Xbar) dimensions to 256x256. In the second implementation, change the crossbar dimensions to 128x128. Report the total latency, power, and energy, and fill in the table. What happened? Why? **(Explain in detail for each parameter.)**

**Table I**

| | 128*128 | 256*256 | Status |
|---|---|---|---|
| Latency | | | Reduce / Rise |
| Power | | | Reduce / Rise |
| Energy | | | Reduce / Rise |

Question 4:

In some layers, the power is the same, while the latency is different. What is the cause?

Question 5:

Which sizing is better? Please explain in detail. For example, if your decision is based on power, latency, or energy, you must mention and explain the reason that causes this condition.