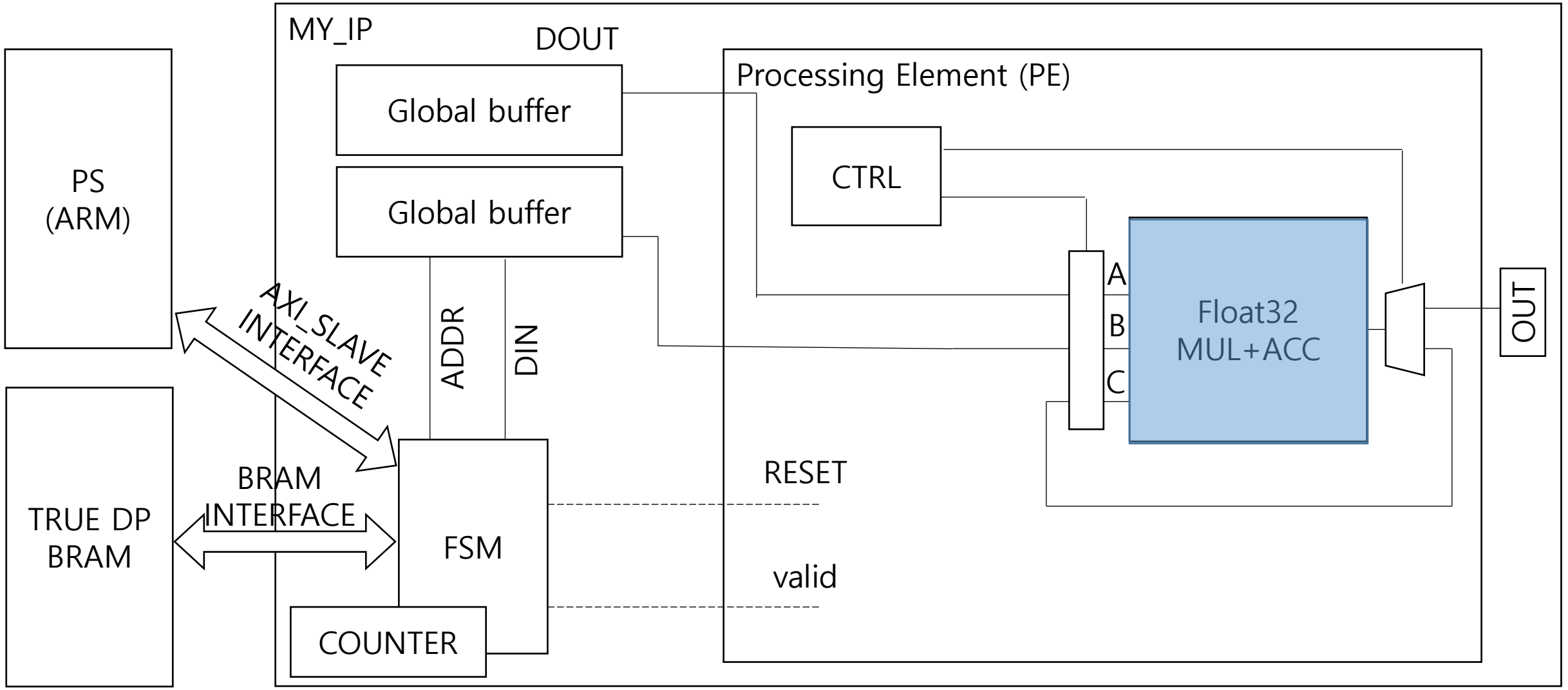
Practice 5

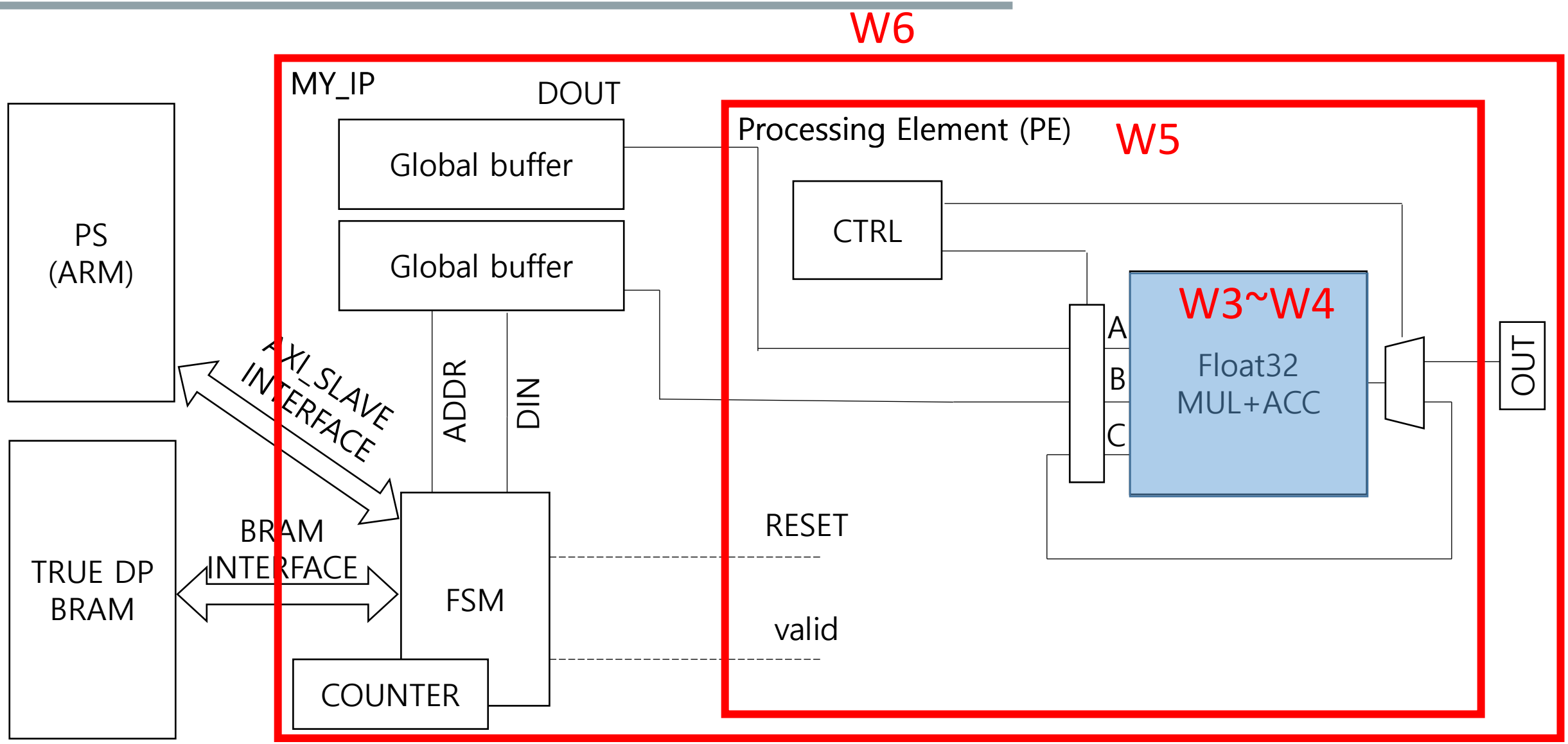
- **PE implementation & BRAM modeling**

Computing Memory Architecture Lab.

Final Project Overview: Matrix Multiplication IP



Final Project Overview: Matrix Multiplication IP



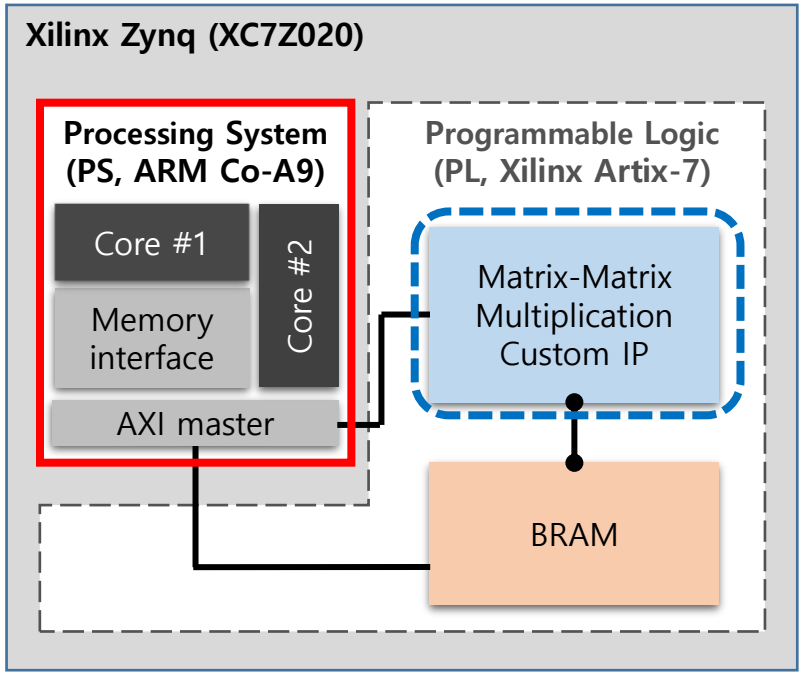
Final Project Overview: Matrix Multiplication IP

- CNN is our application
  - Convolution layer becomes a matrix-matrix multiplication after convolution lowering
    e.g., 32x64 matrix * 64x64 matrix ➔ 32x64 matrix
- ARM CPU runs the main function which calls your MM IP on PL
  - MM for 32x64 weight matrix * 64x64 input matrix multiplication
- BRAM is used for data transfer between SW and HW

MM function on Hardware (Software running on CPU)

```
for(i=0; i<32; i+=1) {
   for(j=0; j<64; j+=1) {
      for(k = 0; k < 64; k++){
         Output[i][j] += Input[i][k]*W[k][j]
      }
   }
}
```

Fused multiply



Main Practice

Practice

1. Implementing BRAM model

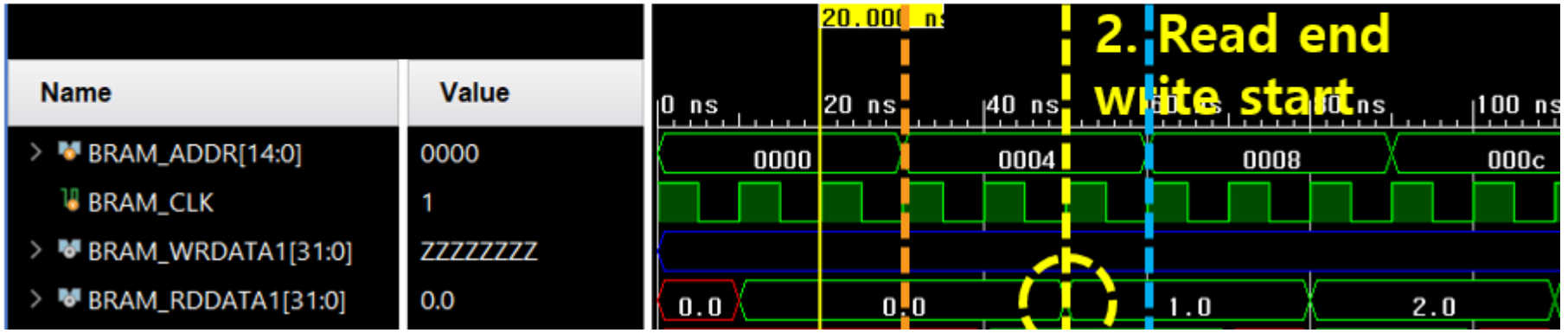
- Implement BRAM model & test-bench according to scenarios below.
  - BRAM costs 1 cycle for write and 2 cycles for read.
  - (i.e., BRAM returns value one cycle after it reads address)

    (= return the data value one cycle after the BRAM recognizes the changed address.)

- Scenario
  - ① **Make test-bench that instantiates two BRAMs and initialize one BRAM to store its address as data.**
    - (i.e., 'mem[0]' stores '0' and 'mem[1]' stores '1')
  - ② **Then, copy every data from the initialized BRAM to the other BRAM.**
    - Use I/O functions of verilog (e.g., '$readmemh' and '$writememh') & text-file (e.g., .txt) to initialize original BRAM and extract values from copied BRAM (access http://sunshowers.tistory.com/10 to get more information about I/O functions of Verilog).

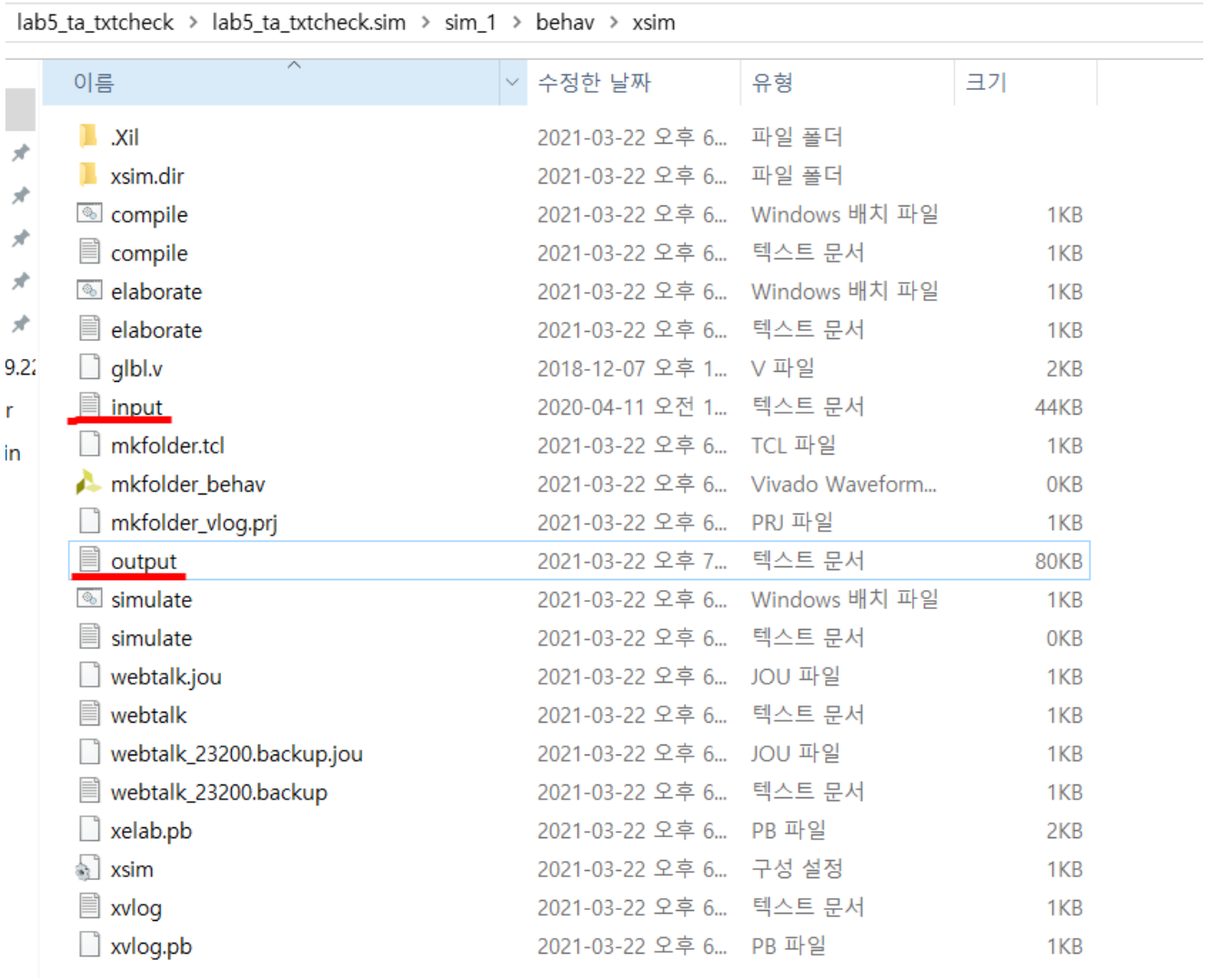
BRAM SIMULATION

- BRAM1 -> BRAM2 operation



- On the waveform, it also take one cycle until BRAM reads address since we assign it,
- So it takes in total two cycles to return value.

Path of File(input.txt, output.txt)



- **Make empty simulation file, and Run simulation to make folder.**

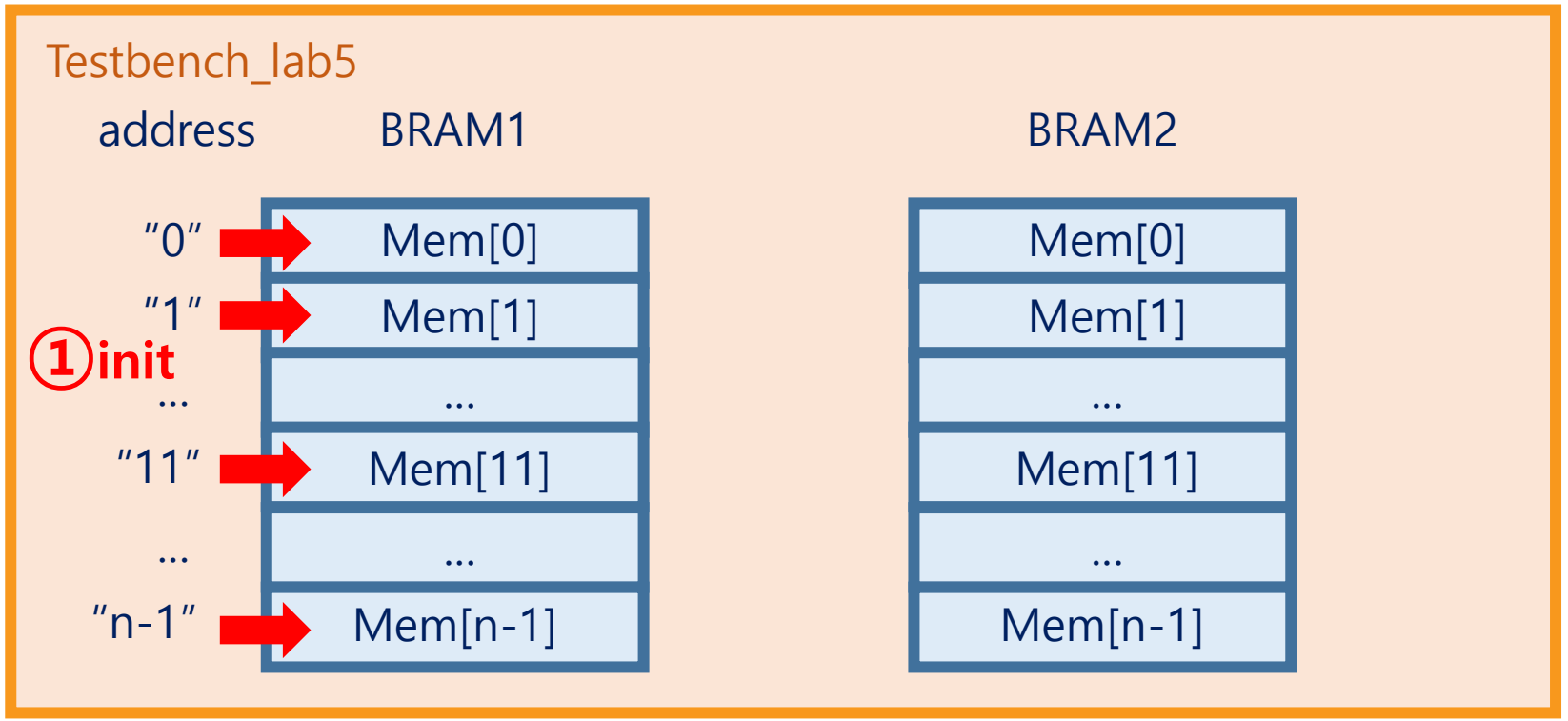
- **From the Project folder, put your input.txt into**

<project_name>/<project_name>.sim/sim_1/behav/xsim/input.txt

Practice

① **Make test-bench that instantiates two BRAMs and initialize one BRAM to store its address as data.**

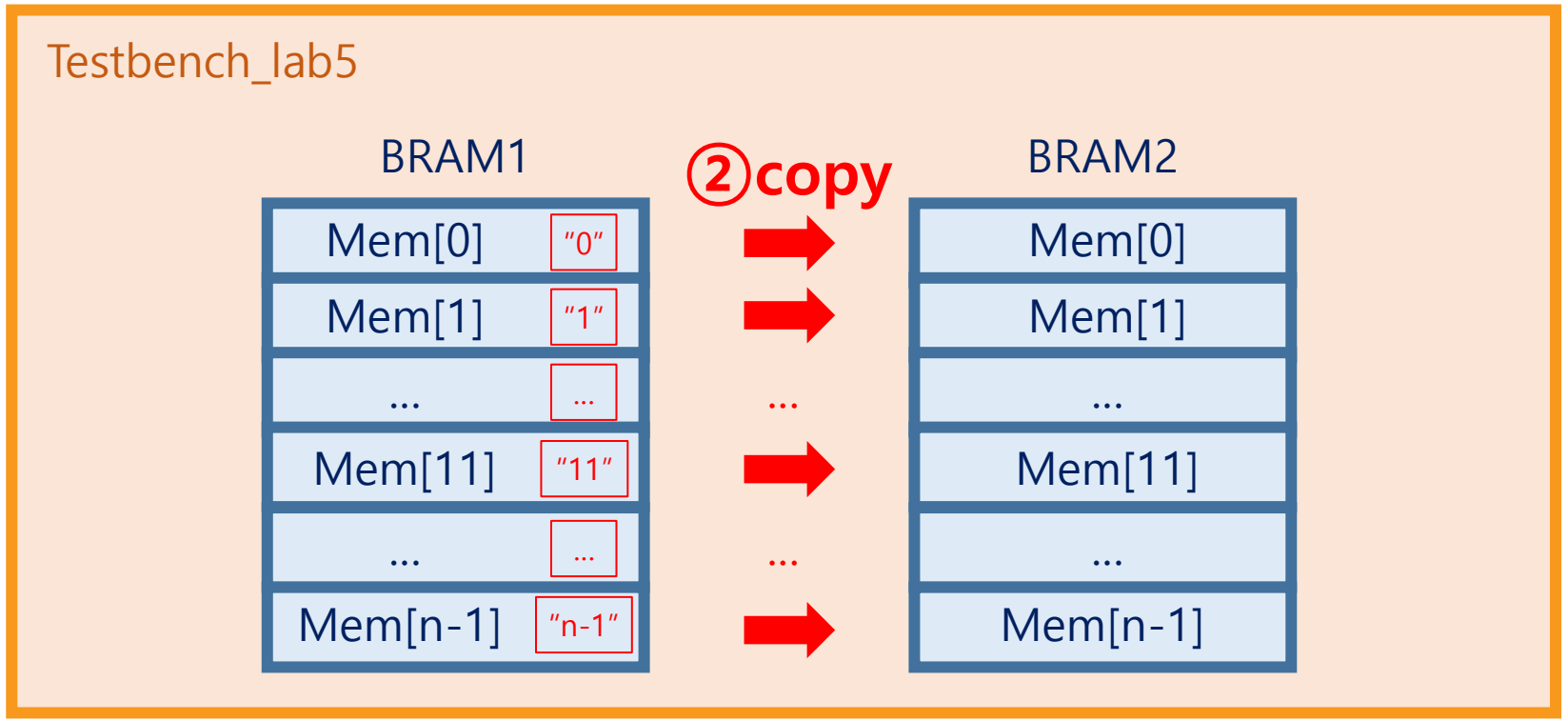
- (i.e., 'mem[0]' stores '0' and 'mem[1]' stores '1')
- Add source -> input.txt



Practice

② **Then, copy every data from the initialized BRAM to the other BRAM.**

- Use I/O functions of verilog (e.g., '$readmemh' and '$writememh') & text-file (e.g., .txt) to initialize original BRAM and extract values from copied BRAM



Practice

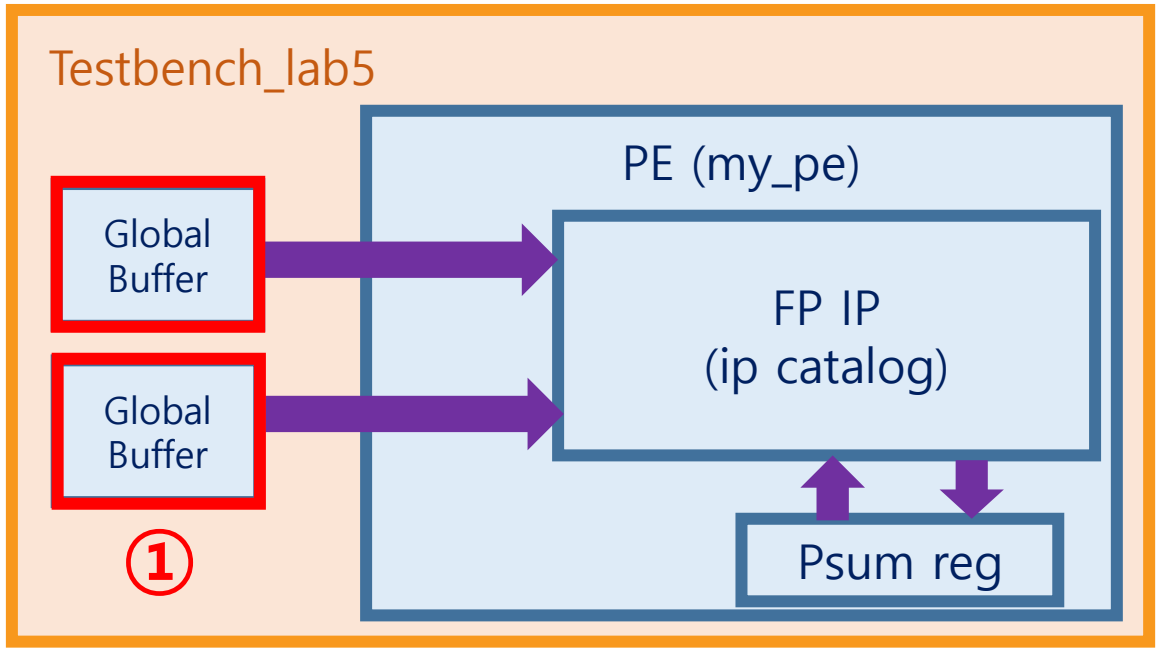
2. Implementing PE with floating-point fused multiply-adder

- Implement Processing Element (PE) & test-bench according to scenarios below.
- Processing Element (PE) consists of floating-point fused multiply-adder constructed with IP catalog (i.e., you implemented last week).

- Scenario
  - ① **There are two global buffer outside of PE, and each memory stores 16 data which is already defined with consecutive addressed (you can set it in TestBench)**
  - ② **Then, PE gets 16 new data from register outside (test-bench) serially to perform MAC (Multiply-Accumulate) operations with the data stored in local register.**
    - Fused multiply-add: $result = (a_{in} * b_{in}) + c_{in}$
    - MAC (Multiply-Accumulate): $result = (a_{in} * b_{in}) + result$
  - ③ **Check 16 MAC results in testbench.**
    - Note that one MAC operation costs several cycles to compute result and it sets valid bit (i.e. dvalid) high when it finishes each operations.

PE schematic

① **There are two global buffers outside of PE, and each global buffer stores 16 data which is already defined with consecutive addressed (you can set it in TestBench)**

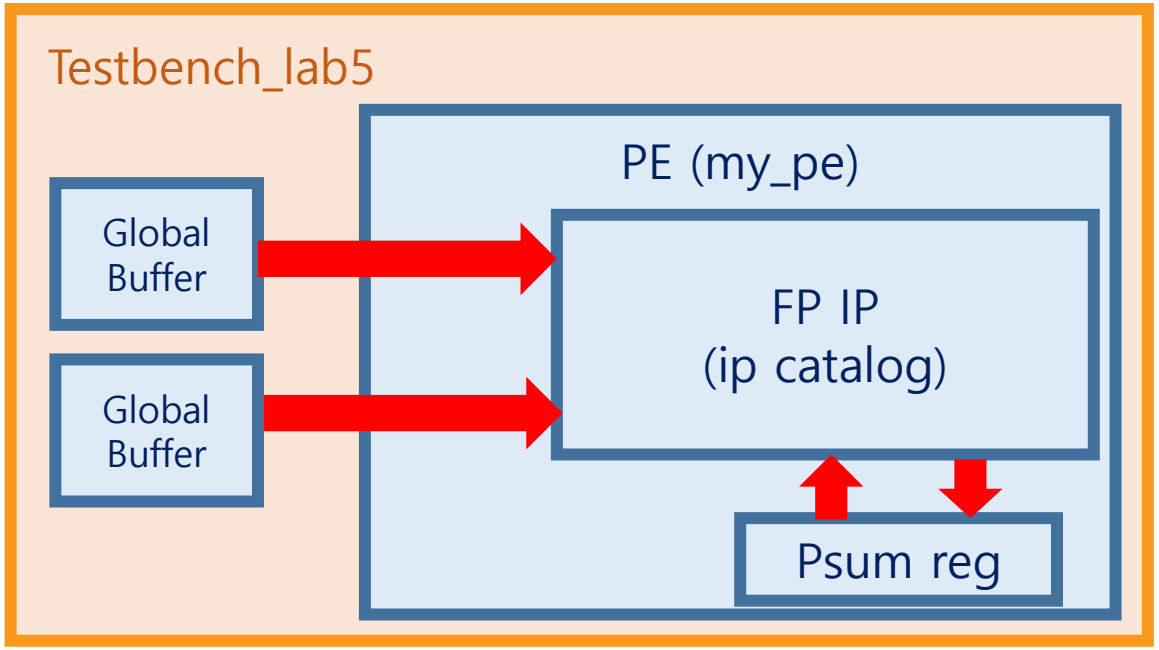


psum <= psum + ain * bin

PE schematic

② **Then, PE gets each 16 data from global buffer outside serially to perform MAC operations with the data stored in local register.**

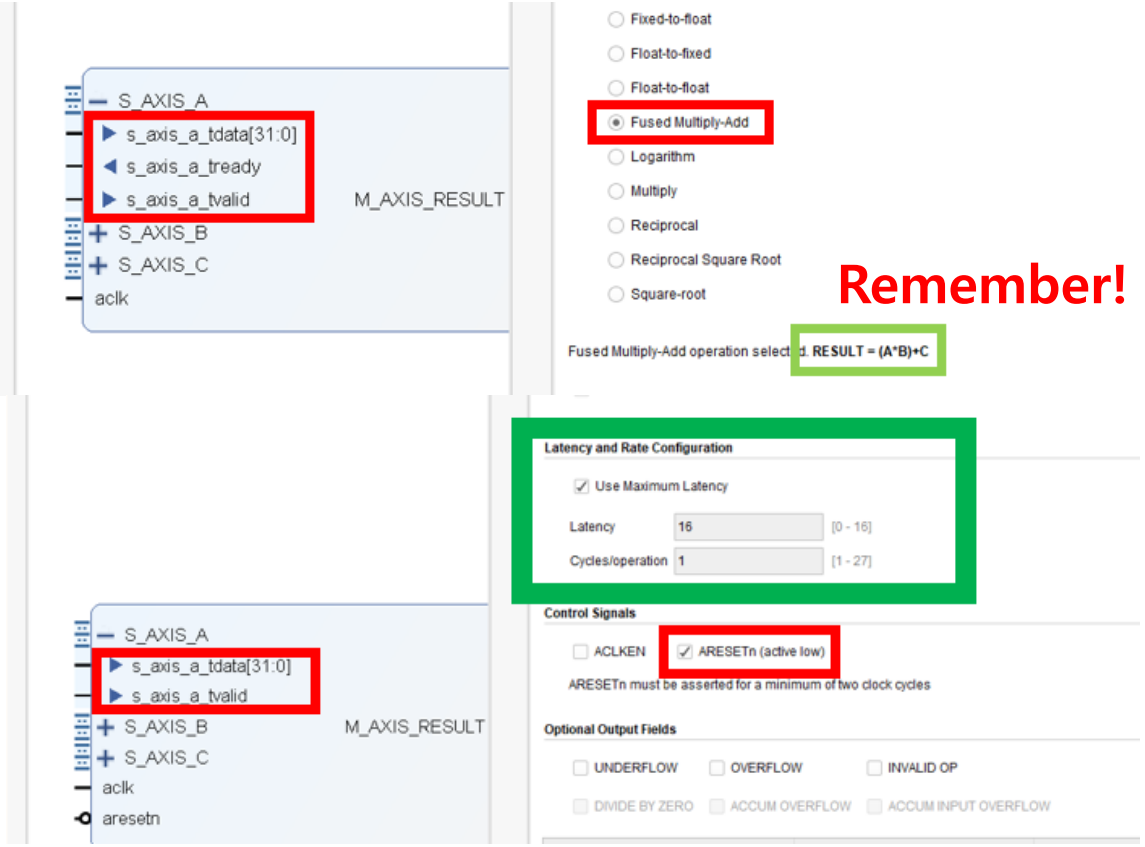
- MAC (Multiply-Accumulate): $result = (a_{in} * b_{in}) + result$

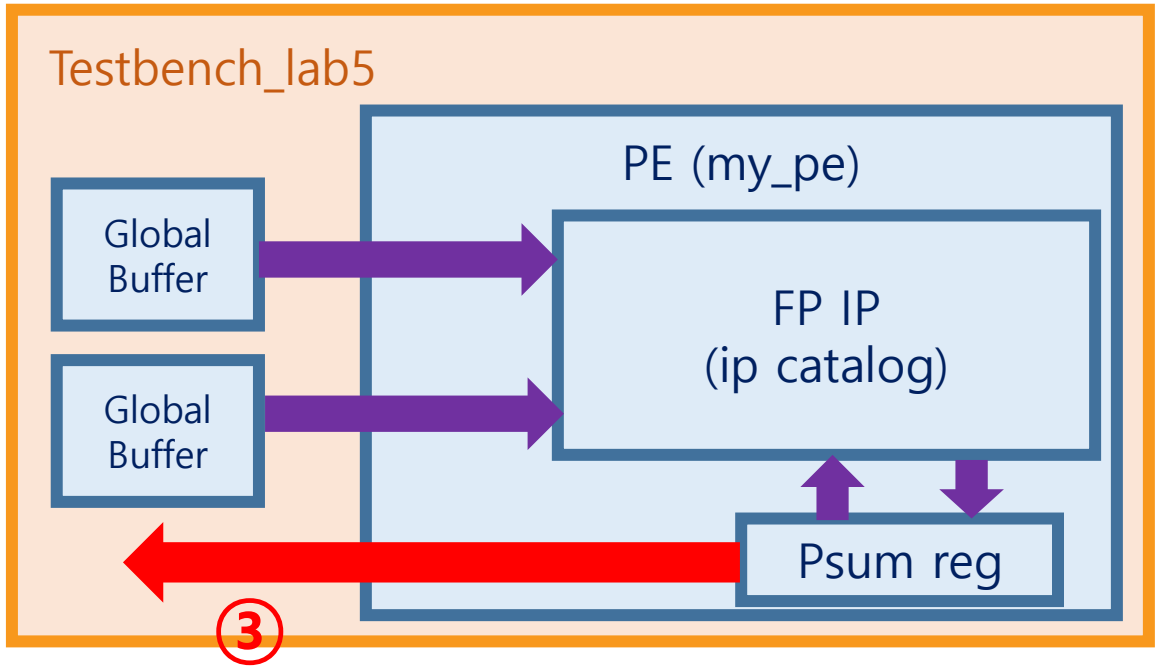


PE schematic

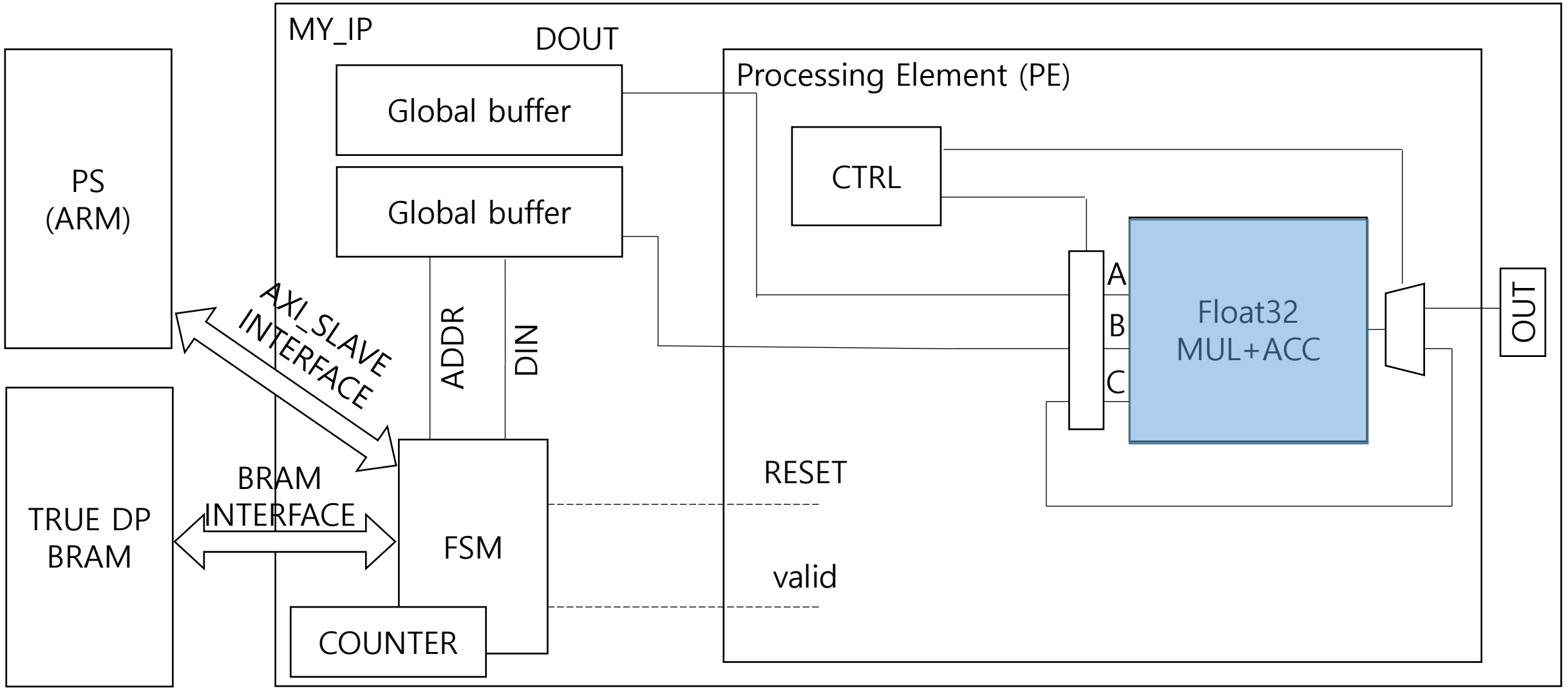
③ Check 16 MAC results.

- Note that one MAC operation costs several cycles to compute result and it sets valid bit (i.e. dvalid) high when it finishes each operations.





Final Project Overview: Matrix Multiplication IP



BRAM Model

```
module my_bram # (
    parameter integer BRAM_ADDR_WIDTH = 15, // 4x8192
    parameter INIT_FILE = "input.txt",
    parameter OUT_FILE = "output.txt"
  )(
    input wire [BRAM_ADDR_WIDTH-1:0] BRAM_ADDR,
    input wire BRAM_CLK,
    input wire [31:0] BRAM_WRDATA,
    output reg [31:0] BRAM_RDDATA,
    input wire BRAM_EN,
    input wire BRAM_RST,
    input wire [3:0] BRAM_WE,
    input wire done
  );
    reg [31:0] mem[0:8191];
    wire [BRAM_ADDR_WIDTH-3:0] addr = BRAM_ADDR[BRAM_ADDR_WIDTH-1:2];
    reg [31:0] dout;

    // code for reading & writing
    initial begin
      if (INIT_FILE != "") begin
          // read data from INIT_FILE and store them into mem
          ...
      end
      wait (done)
          // write data stored in mem into OUT_FILE
          ...
    end

   //code for BRAM implementation
    ...

endmodule
```

- **Memory configuration**
  - **Bit width of data:** 32 bits (4 bytes)
  - **Bit width of address:** 15 bits
  - **# of memory entries:** 8192 (== 2^{13})
  - Each external address of which size is 15 bits (i.e., **BRAM_ADDR**) is associated with 1 byte of data. Last two bits are masked and assigned to internal address of which size is 13 bits (i.e., **addr**). It is associated with each entry of **mem** (i.e. it is associated with 4 bytes of data).
- **BRAM_ADDR:** external address
- **BRAM_CLK:** clock signal
- **BRAM_WRDATA:** Data input port of BRAM.
- **BRAM_RDDATA:** Data output port of BRAM.
- **BRAM_EN:**
  - **(BRAM_EN==1): BRAM enabled. BRAM is available for read or write operation.**
  - If all of **BRAM_WE** is equal to 0 and **BRAM_EN** == 1, read **mem[addr]** into **BRAM_RDDATA**
- **BRAM_WE:**
  - (**BRAM_WE[i]**==1): **BRAM_WRDATA[8*(i+1)-1:8*i]** is stored into **mem[addr][8*(i+1)-1:8*i].**
- **BRAM_RST:** ==1, **BRAM_RDDATA** prints 0
- **done:** ==1, write data stored in mem into **"OUT_FILE"**

Processing Element (PE)

```
module my_pe #(
        parameter L_RAM_SIZE = 6
    )
    (
        // clk/reset
        input aclk,
        input aresetn,
        // port A
        input [31:0] ain,
        // port B
        input [31:0] bin,
        // integrated valid signal
        input valid
        // computation result
        output dvalid,
        output [31:0] dout
    );

    ...

endmodule
```

- **aclk:** clock signal
- **aresetn:** negative reset; ==0, reset is activated;

  Do not forget to reset in first time in testbench.
- **ain:** input ports connected to MAC
- **bin:** input ports connected to MAC
- **valid:** ==1, MAC gets inputs from its input ports and starts computation; It is divided into three valid signals and assigned to MAC (i.e., it is assigned to **s_axis_a_tvalid, s_axis_b_tvalid** and **s_axis_c_tvalid**).
- **dvalid:**
  - ==1, result data from MAC is valid;
  - ==0, result data from MAC is not valid; It is assigned to result valid signal of MAC (i.e., **m_axis_result_tvalid**).
- **dout:** if(dvalid==1), it prints result data from MAC; otherwise it prints 0.

Processing Element (PE) Testbench

```
module tb_my_pe #(
        parameter L_RAM_SIZE = 6
    )

   // global buffer
   reg [31:0] gb1 [0:2**L_RAM_SIZE - 1];
   reg [31:0] gb2 [0:2**L_RAM_SIZE - 1];

   initial begin
        ...
   end

   ...

endmodule
```

Homework

- Requirements
  - Result
    - Attach your project folder with all your verilog codes (e.g., BRAM, PE, test bench)
    - Attach your BRAM waveform(simulation result) with [student_number, name]
      - Test the scenario in slide6.
      - The correct waveform should be shown to confirm the operation of your code.
      - Refer to Practice3 about screenshot.
    - Attach your PE waveform(simulation result) with [student_number, name]
      - Test the scenario in slide9.
      - The correct waveform should be shown to confirm the operation of your code.
      - Refer to Practice3 about screenshot.
  - Report
    - Explain operation of BRAM with waveform that you implemented
    - Explain operation of PE with waveform that you implemented
    - In your own words
    - Either in Korean or in English
    - # of pages does not matter
    - **PDF only!!**
  - **Result + Report to one .zip file**
- Upload (.zip) file on ETL
  - Submit one (.zip) file
    - zip file name : [Lab05]name1_name2.zip (ex : [Lab05] 홍길동 홍동길 .zip --> **All Team members' name should be included**)
  - Due: 4/7(WED) 23:59, Start Early!
    - **No Late Submission**