

12 BONUS: Systolic Arrays [35 points]

A systolic array consists of 4x4 Processing Elements (PEs), interconnected as shown in Figure 1. The inputs of the systolic array are labeled as H_0, H_1, H_2, H_3 and V_0, V_1, V_2, V_3 . Figure 2 shows the PE logic, which performs a multiply and accumulate MAC operation and saves the result to an internal register (*reg*). Figure 2 also shows how each PE propagates its inputs. We make the following assumptions:

- The latency of each MAC operation is one cycle.
- The propagation of the values from i_0 to o_0 , and from i_1 to o_1 , takes one cycle.
- The initial values of all internal registers is zero.

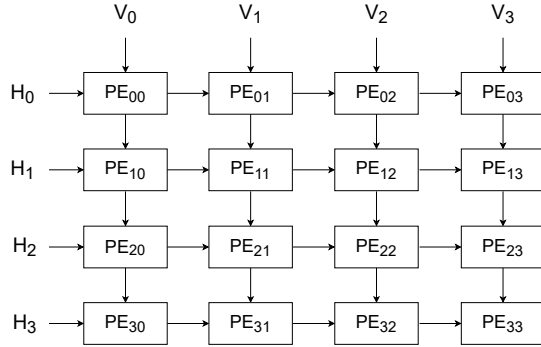


Figure 1: PE array

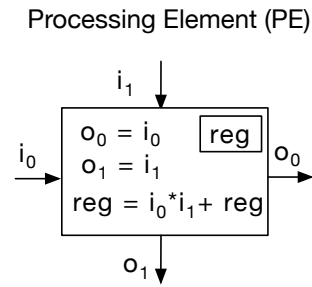


Figure 2: Processing Element (PE)

Your goal is to use the example systolic array shown in Figure 1 to perform the convolution (\otimes) of a 3x3 image (matrix $I_{3 \times 3}$) with four 2x2 filters (matrices $A_{2 \times 2}$, $B_{2 \times 2}$, $C_{2 \times 2}$, and $D_{2 \times 2}$), to obtain four 2x2 outputs (matrices $W_{2 \times 2}$, $X_{2 \times 2}$, $Y_{2 \times 2}$, and $Z_{2 \times 2}$):

$$\begin{matrix} I_{00} & I_{01} & I_{02} \\ I_{10} & I_{11} & I_{12} \\ I_{20} & I_{21} & I_{22} \end{matrix} \quad (\otimes) \quad \begin{matrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{matrix} = \begin{matrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{matrix}$$

$$\begin{matrix} I_{00} & I_{01} & I_{02} \\ I_{10} & I_{11} & I_{12} \\ I_{20} & I_{21} & I_{22} \end{matrix} \quad (\otimes) \quad \begin{matrix} B_{00} & B_{01} \\ B_{10} & B_{11} \end{matrix} = \begin{matrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{matrix}$$

$$\begin{matrix} I_{00} & I_{01} & I_{02} \\ I_{10} & I_{11} & I_{12} \\ I_{20} & I_{21} & I_{22} \end{matrix} \quad (\otimes) \quad \begin{matrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{matrix} = \begin{matrix} Y_{00} & Y_{01} \\ Y_{10} & Y_{11} \end{matrix}$$

$$\begin{matrix} I_{00} & I_{01} & I_{02} \\ I_{10} & I_{11} & I_{12} \\ I_{20} & I_{21} & I_{22} \end{matrix} \quad (\otimes) \quad \begin{matrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{matrix} = \begin{matrix} Z_{00} & Z_{01} \\ Z_{10} & Z_{11} \end{matrix}$$

As an example, the convolution of the matrix $I_{3 \times 3}$ with the filter $A_{2 \times 2}$ is computed as follows:

- $W_{00} = I_{00} * A_{00} + I_{01} * A_{01} + I_{10} * A_{10} + I_{11} * A_{11}$
- $W_{01} = I_{01} * A_{00} + I_{02} * A_{01} + I_{11} * A_{10} + I_{12} * A_{11}$
- $W_{10} = I_{10} * A_{00} + I_{11} * A_{01} + I_{20} * A_{10} + I_{21} * A_{11}$
- $W_{11} = I_{11} * A_{00} + I_{12} * A_{01} + I_{21} * A_{10} + I_{22} * A_{11}$

You should compute the four convolutions in the minimum possible number of cycles. Fill the following table with:

1. The input elements (from matrices $I_{3 \times 3}$, $A_{2 \times 2}$, $B_{2 \times 2}$, $C_{2 \times 2}$, and $D_{2 \times 2}$) in the correct input ports of the systolic array (H_0 , H_1 , H_2 , H_3 and V_0 , V_1 , V_2 , V_3). (Hint: If necessary, an input element can be concurrently streamed into several input ports of the array.)
2. The output values and the corresponding PE where the output elements (of matrices $W_{2 \times 2}$, $X_{2 \times 2}$, $Y_{2 \times 2}$, and $Z_{2 \times 2}$) are generated.

Fill the blanks only with relevant information.

cycle	H0	H1	H2	H3	V0	V1	V2	V3	PE ₀₀	PE ₀₁	PE ₀₂	PE ₀₃	PE ₁₀	PE ₁₁	PE ₁₂	PE ₁₃	PE ₂₀	PE ₂₁	PE ₂₂	PE ₂₃	PE ₃₀	PE ₃₁	PE ₃₂	PE ₃₃
0	A ₀₀				I ₀₀																			
1	A ₀₁	B ₀₀			I ₀₁	I ₀₁																		
2	A ₁₀	B ₀₁	C ₀₀		I ₁₀	I ₀₂	I ₁₀																	
3	A ₁₁	B ₁₀	C ₀₁	D ₀₀	I ₁₁	I ₁₁	I ₁₁	I ₁₁	W ₀₀															
4		B ₁₁	C ₁₀	D ₀₁		I ₁₂	I ₂₀	I ₁₂		W ₀₁			X ₀₀											
5			C ₁₁	D ₁₀			I ₂₁	I ₂₁			W ₁₀			X ₀₁			Y ₀₀							
6				D ₁₁				I ₂₂				W ₁₁			X ₁₀			Y ₀₁			Z ₀₀			
7																X ₁₁			Y ₁₀			Z ₀₁		
8																				Y ₁₁			Z ₁₀	
9																								Z ₁₁
10																								
11																								
12																								
13																								
14																								
15																								