Digital IC Design

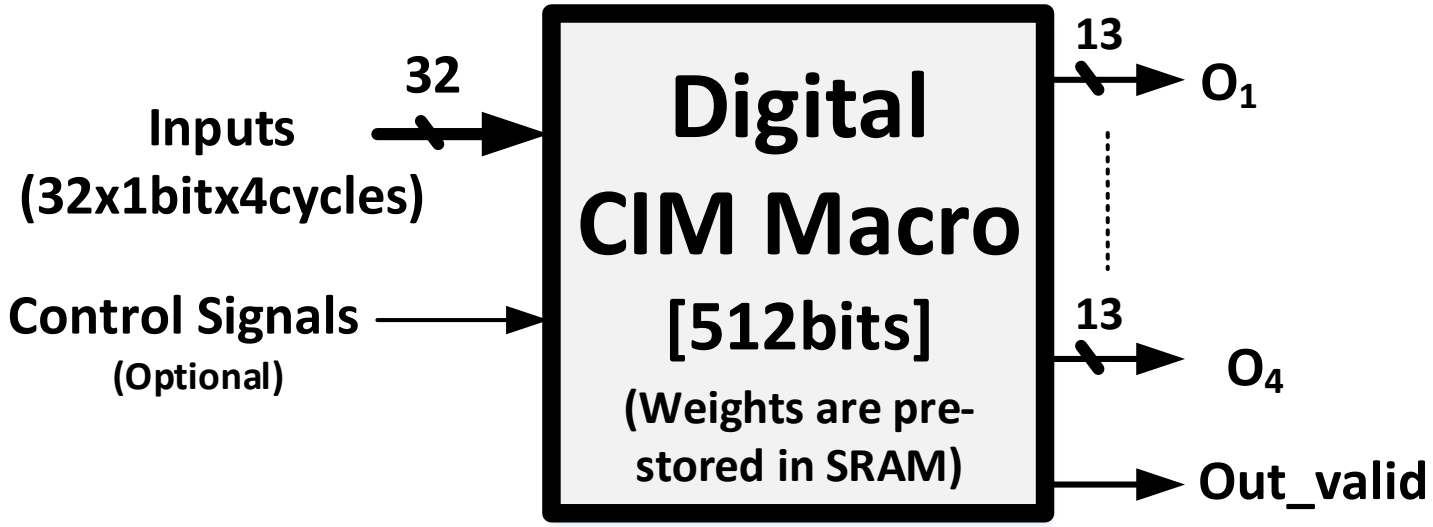
Final Team Project:
32-Rows x 4-Columns Digital Computation-in-Memory Macro for Matrix Multiplications

Professor Po-Tsang Huang

International College of Semiconductor Technology
National Yang Ming Chiao Tung University

Design a Digital CIM Macro

- Spec of CIM macro:
  - Array Size:
    - 512bit
    - 32rows x 4 columns x 4bit
  - CIM cell:
    - 6T SRAM + NOR (Provided by TA)
  - Pre-store data:
    - Weights should be stored in SRAM cells by initial conditions (.IC)



Design a Digital CIM Macro

- Spec of matrixes:
  - Inputs matrix (unsigned values):
    - 4 rows x 32columns: $(I_1, I_2, \ldots\ldots I_{32})_1$, $(I_1, I_2, \ldots\ldots I_{32})_2$, … $(I_1, I_2, \ldots\ldots I_{32})_4$
    - All inputs are 4bit
    - **Give the input bit serially:** 32 inputs (4 bits each) should be sent to the DCIM macro, 1 bit per cycle, over a total of 4 cycles to generate a column of output
    - You can add any extra control signals for your design
  - Weights matrix(unsigned values):
    - 32rows x 4 columns: $(W_1, W_2, W_3, W_4)_1$, $(W_1, W_2, W_3, W_4)_2$, … $(W_1, W_2, W_3, W_4)_{32}$
    - All weights are 4bit
    - **Weights should be stored in latches by initial conditions** (.IC)
  - Output matrix (unsigned values):
    - 4rows x 4columns: $(O_1, O_2, O_3, O_4)_1$, $(O_1, O_2, O_3, O_4)_2$, … $(O_1, O_2, O_3, O_4)_4$,
    - All outputs are 13 bits
    - $(O_1)_1 = ((I_1)_1 \times (W_1)_1) + ((I_2)_1 \times (W_1)_2) + \ldots ((I_{32})_1 \times (W_1)_{32})$
    - $(O_2)_1 = ((I_1)_1 \times (W_2)_1) + ((I_2)_1 \times (W_2)_2) + \ldots ((I_{32})_1 \times (W_2)_{32})$
    - $(O_1)_2 = ((I_1)_2 \times (W_1)_1) + ((I_2)_2 \times (W_1)_2) + \ldots ((I_{32})_2 \times (W_1)_{32})$

Design a Digital CIM Macro

- Input & Output ports

Input signal	Bit width	Definition
I_1-I_{32}	1bit x 32	Input serial signals
Output signal	Bit width	Definition
Out_valid	1	O_1- O_4 are valid
O_1 – O_4	13bit x 4	Output signals

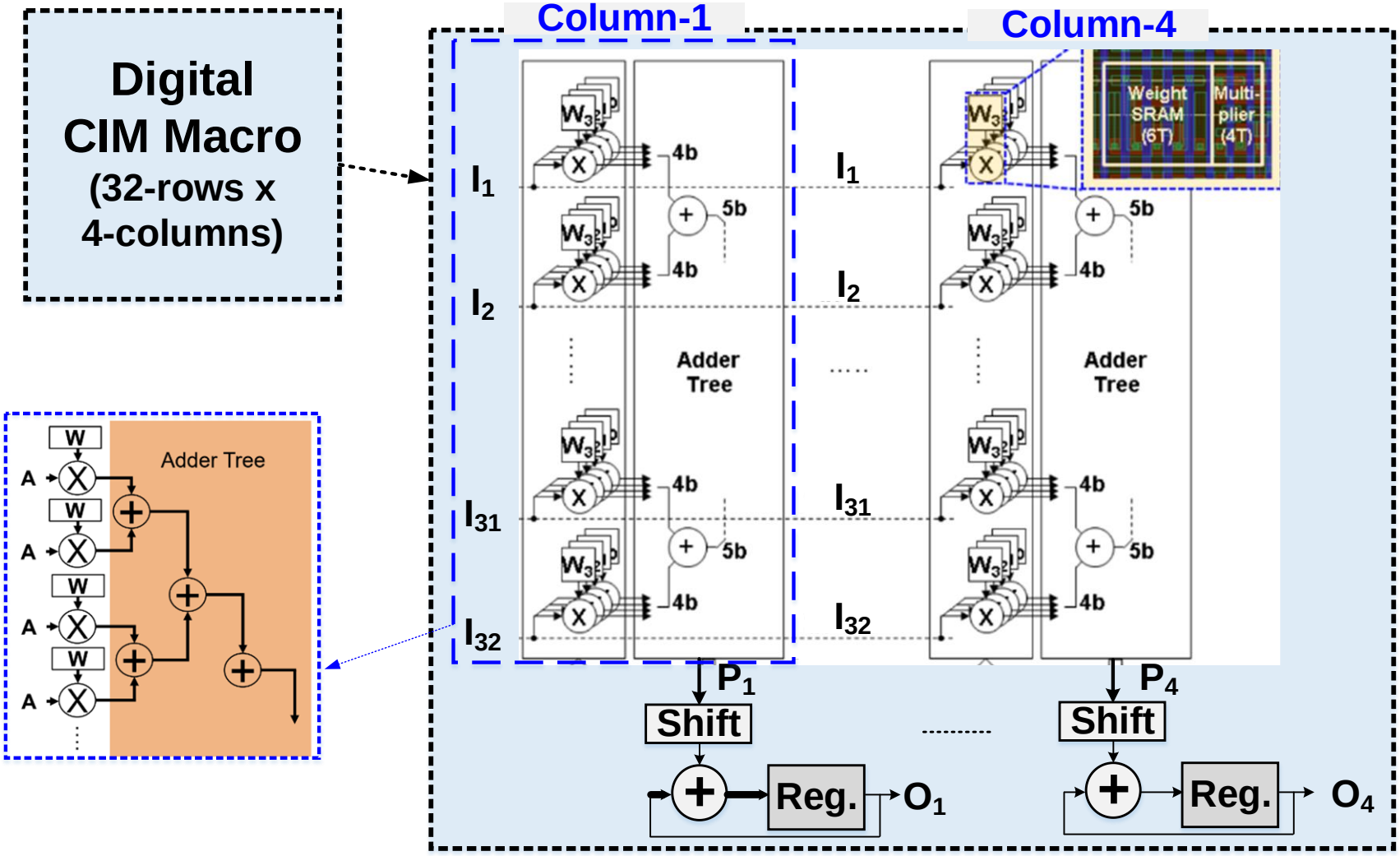
*You can add Control signals and define by yourself

- Matrixes:

$$\begin{bmatrix} (I_1)_1 & (I_2)_1 & \cdots & (I_{32})_1 \\ (I_1)_2 & (I_2)_2 & \cdots & (I_{32})_2 \\ \vdots & \ddots & & \\ (I_1)_4 & (I_2)_4 & & (I_{32})_4 \end{bmatrix} \times \begin{bmatrix} (W_1)_1 & (W_2)_1 & \cdots & (W_4)_1 \\ (W_1)_2 & (W_2)_2 & \cdots & (W_4)_2 \\ \vdots & \ddots & & \\ (W_1)_{32} & (W_2)_{32} & & (W_4)_{32} \end{bmatrix} = \begin{bmatrix} (O_1)_1 & (O_2)_1 & \cdots & (O_4)_1 \\ (O_1)_2 & (O_2)_2 & \cdots & (O_4)_2 \\ \vdots & \ddots & & \\ (O_1)_4 & (O_2)_4 & & (O_4)_4 \end{bmatrix}$$

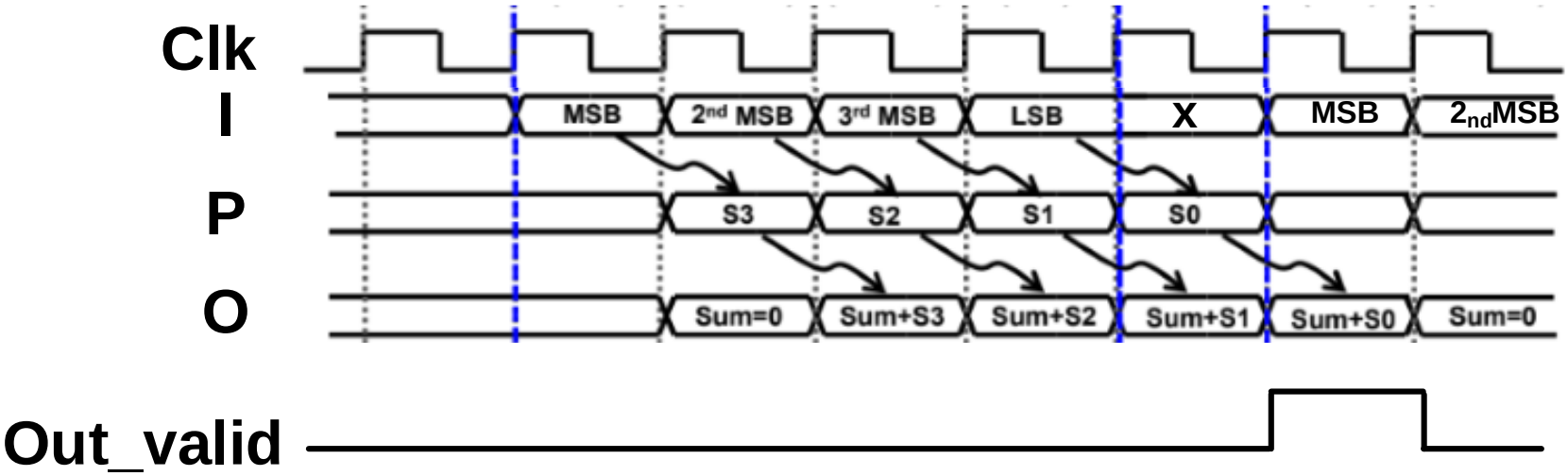
Digital CIM Macro

- Example of 32-rows x 4-columns DCIM macro



Example of Bit Serial Input

- Example of Waveform
  - Bit Serial input



Maximize the energy efficiency

- Maximize the energy efficiency **under 2 different throughput, respectively:**
  - When the **throughput** is higher than **30 GOPS**
  - When the **throughput** is higher than **80 GOPS**
  - 1MAC = 2 OPs
- The Frequency and Voltage can be adjusted by yourself
- The input/weight matrixes should be random patterns
- The function should be correct
  - Provide correct waveform (Each Outputs should be combined into a bus signal(13bit) in Decimal format)
  - Example of 13bit bus signal in Decimal format :